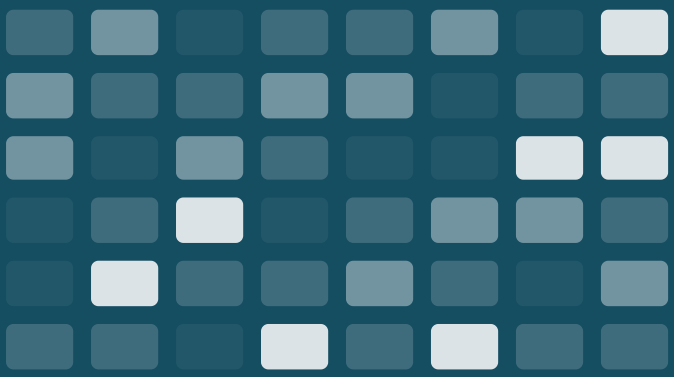
Иван Бегтин



Каталоги данных

Как устроены порталы открытых данных,
геопорталы и репозитории — и как выбрать свой

Редакция октября 2026 года

Каталоги данных. Как устроены порталы открытых данных, геопорталы и репозитории — и как выбрать свой

© 2026 Иван Бегтин. Редакция октября 2026 года.

Текст доступен по лицензии Creative Commons «Атрибуция» 4.0 Международная (CC BY 4.0): creativecommons.org/licenses/by/4.0/deed.ru.



Книга и база знаний о каталогах данных создаются при поддержке Dateno — поисковой системы по наборам данных.

Онлайн-версия книги: https://datacatalogs.begtin.tech/

Исходный текст: https://github.com/ivbeg/data-catalogs-book

PDF собран автоматически 23 октября 2026 г., версия исходного текста 2b56130. Самая свежая редакция и интерактивная маршрутная карта — на сайте книги.

Содержание

О книге 6

I. Что такое каталог данных 12

Что такое каталог данных 13

Зачем создают каталоги и кто ими пользуется 16

Анатомия каталога 19

Мир каталогов в цифрах 27

II. Какие данные публикуют 31

Что публикуют в каталогах 32

Единицы публикации и форматы 40

III. Типы каталогов 45

Как классифицировать каталоги 46

Порталы открытых данных 57

Геопорталы 67

Научные репозитории данных 80

Каталоги показателей 93

Каталоги микроданных 106

ML-каталоги 119

Каталоги метаданных и реестры 131

Поисковики и агрегаторы 145

Каталоги API, маркетплейсы и списки наборов 159

Сравнение типов и гибриды 172

IV. Возможности каталогов 184

Карта возможностей 185

Экспорт метаданных 195

Полнота и детальность метаданных 205

API и массовая выгрузка 217

Внутреннее хранилище 231

Постоянные идентификаторы, DOI и версии 243

Экспорт цитирования 257

Доступ, роли и процессы 269

Скорость развёртывания и сопровождение 286

Другие возможности 290

V. Платформы и расширения 300

Классы программного обеспечения 301

Экосистемы расширений 304
CKAN: ядро и расширения 311
Карточки платформ 315

VI. Как выбрать каталог 333
Маршрутная карта выбора 334
Принципы выбора набора функций 339
От функций к платформе 347
Как обосновать решение 351
Свои силы, свободное ПО или облачный сервис 358
Типовые сценарии 365
Чек-лист требований 374
Разобранные примеры 377

VII. Как публикуют данные 384
Процесс публикации 385
Стандарты метаданных 392
Федерации и агрегация 401
Лицензии и условия доступа 406
Находимость 415

VIII. Публикация без портала 420
Когда портал не нужен 421
Статические сайты 428
Git-репозитории 436
Облачные хранилища и общие хабы 444
Как сделать публикацию находимой 452

IX. Практики 460
Хорошие практики 461
Антипаттерны 467
Жизненный цикл каталога 472
Архивация и спасение каталогов 478

X. Качество каталогов 483
Уровни качества 484
Метрики качества 489
Методики и рейтинги 495
Самоаудит каталога 502

XI. Тренды 509
ИИ, пространства данных и регулирование 510

Что уже умеют передовые каталоги 520

Приложения **529**

Глоссарий 530

Как работать с реестром 548

Сравнительные таблицы платформ 572

Чек-листы 577

Литература и ссылки 585

Стабильные адреса для ссылок 590

Об авторе **591**

О книге

Как устроены порталы открытых данных, геопорталы и репозитории — и как выбрать свой.





Автор — Иван Бегтин, основатель и директор АНО «Информационная культура», основатель и технический директор поисковика по наборам данных Dateno. Открытыми данными занимается с 2009 года, каталогами — с 2010-го: от общественного «Хаба открытых данных» до реестра каталогов всего мира, по которому посчитаны цифры этой книги. Подробнее — в главе «Об авторе».

Книга и база знаний о каталогах данных создаются при поддержке Dateno — поисковой системы по наборам данных. Dateno собирает описания наборов из тысяч каталогов по всему миру: порталов открытых данных, геопорталов, научных репозиториев и статистических служб — и ищет по ним в одном месте. На 16 октября 2026 года в нём 22 миллиона наборов из 5 236 источников.

Редакция октября 2026 года. Текст доступен по лицензии CC BY 4.0.

Скачать книгу в PDF: на русском · in English. Английская версия сайта — в переключателе языков в меню.

Каталог данных — сервис, который описывает наборы данных и помогает их найти, понять и получить. Эта книга — о публичных каталогах: порталах открытых данных, геопорталах, научных репозиториях, каталогах статистики и наборов для машинного обучения. Она объясняет, какие каталоги бывают и что они умеют, как выбрать функции и платформу и обосновать выбор, как публиковать данные — на портале или без него.

Корпоративные каталоги, которые описывают данные внутри компании (Collibra, Alation, DataHub), в книге упоминаются только для сравнения.

Если вы ищете не каталог, а сами данные, начните с поиска в Dateno. Как он связан с цифрами книги — в разделе «Откуда цифры».

С чего начать

Книгу не нужно читать подряд. Выберите свою задачу: под ней 4–6 глав в порядке чтения и то, что будет у вас в конце. Если ни одна задача не подходит, пройдите маршрутную карту: восемь вопросов, около пяти минут.

<details open>
<summary>**Поручили запустить портал открытых данных** — орган власти, регион, город · 6 глав, около 1 ч 15 мин</summary>

1. Маршрутная карта выбора — первый ориентир по типу каталога и платформе.
2. Типовые сценарии — найдите свою строку: центральный орган, регион или муниципалитет.
3. Когда портал не нужен — может быть, задачу уже закрывает региональный или национальный портал.
4. Принципы выбора набора функций — список функций из задач, а не из рекламы.
5. Чек-лист требований — что положить в техническое задание.
6. Как обосновать решение — карточка решения для руководителя и для следующего сотрудника.

В конце у вас: ответ, нужен ли свой портал, список функций для технического задания и карточка решения.
</details>

<details>
<summary>**Выбрать платформу или написать техническое задание** · 6 глав, около 1 ч 45 мин</summary>

1. Карта возможностей — какие возможности бывают и что типично для вашего типа каталога.
2. Принципы выбора набора функций — у каждой функции пять пометок: чья задача, насколько обязательна, почему, как проверить, когда пересмотреть.
3. От функций к платформе — как превратить список функций в короткий список платформ.
4. Экосистемы расширений — как проверить расширение до того, как платформа попадёт в список.
5. Карточки платформ — что у каждой платформы в ядре, что в расширениях, какие ограничения.
6. Чек-лист требований.

В конце у вас: короткий список платформ с пометками «в ядре», «в расширении», «нет» и требования для технического задания.
</details>

<details>
<summary>**Опубликовать данные к статье, нужен DOI** · 5 глав, около 1 ч 35 мин</summary>

1. Научные репозитории данных — Zenodo, профильный или университетский репозиторий.
2. Постоянные идентификаторы, DOI и версии — чем DOI версии отличается от общего DOI.
3. Экспорт цитирования — как набор будут цитировать и зачем файл `CITATION.cff`.
</details>

4. Лицензии и условия доступа — какую лицензию поставить и что она разрешает.
5. Как сделать публикацию находимой — чтобы набор нашли поисковики и агрегаторы.

В конце у вас: выбранный репозиторий, DOI, лицензия и файл цитирования.

Опубликовать несколько наборов без своего портала · 4 главы, около 50 мин

1. Когда портал не нужен — пять вопросов до установки.
2. Статические сайты или Git-репозитории — выберите один способ.
3. Как сделать публикацию находимой — страница набора, файл описаний и ссылки из чужих каталогов.
4. Чек-лист «Минимум публикации без портала» и шаблоны файлов — README, `datapackage.json`, `CITATION.cff`, разметка schema.org.

В конце у вас: страница или репозиторий с описаниями наборов, которые находят поисковики и агрегаторы.

Собирать описания из многих каталогов — разработчик поисковика или агрегатора · 5 глав, около 1 ч 50 мин

1. Поисковики и агрегаторы данных — чем поисковик отличается от агрегатора.
2. Экспорт метаданных — DCAT, schema.org и что на деле отдают платформы.
3. API и массовая выгрузка — полная выгрузка против поискового API.
4. Федерации и агрегация — как собирают описания и где теряются записи.
5. Стандарты метаданных — какой язык описаний читает каждый адресат.

В конце у вас: какие протоколы и форматы забирать у каждого типа каталогов и какие потери ждать.

Статистическая служба: показатели и микроданные · 5 глав, около 2 ч

1. Разобранный пример статистической службы — почему это два каталога, а не один.
2. Каталоги показателей — временные ряды, SDMX, PxWeb и .Stat Suite.
3. Каталоги микроданных — обезличенные записи, DDI, NADA и доступ по запросу.
4. Карточки платформ — PxWeb, .Stat Suite и NADA.
5. Чек-листы каталога показателей и каталога микроданных.

В конце у вас: два каталога с разными стандартами и короткий список платформ для каждого.

Проверить и улучшить свой каталог · 4 главы, около 45 мин

1. Уровни качества — какой уровень трогает оценка и кто за него отвечает.
2. Самоаудит каталога — как пройти свой каталог за один раз.
3. Антипаттерны — типичные сбои и откуда они берутся.

4. Хорошие практики — шесть вещей, без которых набор нельзя найти или использовать.

Для проверки на живом сайте — чек-лист проверки портала.

В конце у вас: список проблем по уровням качества и порядок починки.

Главное в девяти тезисах

1. **Каталог — это описания, а не склад файлов.** Без описания ссылка на файл — ещё не каталог. → Что такое каталог данных
2. **Тип каталога задаёт то, что в нём публикуют.** От типа зависят стандарт описания и круг подходящих платформ. → Как классифицировать каталоги
3. **Свой портал нужен не всегда.** Пока наборов и публикаторов немного, данные публикуют в чужом каталоге, на своём сайте, в репозитории или на общем хабе. → Когда портал не нужен
4. **Функции выбирают по задачам**, а не по рекламе платформы и не по чужому техническому заданию. → Принципы выбора набора функций
5. **У каждой функции есть адрес:** ядро платформы, названное расширение или нигде. Если обязательная функция держится на расширении, которое не проверяют на вашей версии ядра, платформу вычёркивают. → Экосистемы расширений
6. **Способ создания выбирают раньше марки.** Платформа бесполезна, если её некому обновлять. → Свои силы, свободное ПО или облачный сервис
7. **Решение записывают.** Карточка решения объясняет следующему сотруднику, какое обязательство закрывает платформа и когда выбор пересматривать. → Как обосновать решение
8. **Набор опубликован, когда посторонний может им воспользоваться:** находит карточку, открывает файл, читает лицензию и знает, кто отвечает за данные. → Процесс публикации
9. **У качества четыре уровня.** Высокий балл на одном не исправляет остальные. → Уровни качества

Содержание

Время — оценка для чтения подряд, без примеров кода. Всего около 17 часов.

- **I. Что такое каталог данных** · 4 главы, 30 мин — определение, устройство каталога, мир каталогов в цифрах.
- **II. Какие данные публикуют** · 2 главы, 25 мин — что публикуют, единицы публикации и форматы.
- **III. Типы каталогов** · 11 глав, 4 ч 30 мин — порталы открытых данных, геопорталы, научные репозитории, статистика, ML, метаданные, агрегаторы и гибриды.
- **IV. Возможности каталогов** · 10 глав, 3 ч 40 мин — экспорт метаданных, API, хранилище, DOI, цитирование, доступ и роли.
- **V. Платформы и расширения** · 4 главы, 1 ч 10 мин — классы программ, ядро и расширения, CKAN, карточки платформ.

- **VI. Как выбрать каталог** · 8 глав, 1 ч 40 мин — маршрутная карта, выбор функций и платформы, обоснование, сценарии и примеры.
- **VII. Как публикуют данные** · 5 глав, 1 ч 10 мин — процесс, стандарты метаданных, федерации, лицензии, находимость.
- **VIII. Публикация без портала** · 5 глав, 1 ч — статические сайты, Git, облачные хранилища и хабы.
- **IX. Практики** · 4 главы, 50 мин — хорошие практики, антипаттерны, жизненный цикл, архивация.
- **X. Качество каталогов** · 4 главы, 50 мин — уровни, метрики, методики и рейтинги, самоаудит.
- **XI. Тренды** · 2 главы, 40 мин — ИИ, пространства данных и регулирование; что уже умеют передовые каталоги.
- **Приложения** · справочник — глоссарий, работа с реестром, таблицы платформ, чек-листы, литература.
- **Об авторе** — кто написал книгу и как его проекты связаны с её цифрами.

Как устроены главы

- **Сначала вывод.** Глава начинается с главного вывода, подробности и примеры — ниже.
- **Ядро, расширение или нет.** В главах о возможностях и в карточках платформ для каждой функции сказано, где она живёт: в самой платформе, в отдельном модуле со своим автором и выпусками или нигде.
- **Карточка решения** — одна-две страницы о выборе: контекст, варианты, критерии, решение, почему, когда пересмотреть. Шаблон — в главе «Как обосновать решение».
- **Даты у фактов.** Версии программ и состояние сайтов меняются, поэтому факты записаны «на октябрь 2026 года». Последний раздел каждой главы — «Источники».
- **Незнакомое слово** ищите в глоссарии.

Готовые материалы для работы: чек-листы, сравнительные таблицы платформ, шаблоны файлов и запросы к реестру каталогов.

Откуда цифры

Цифры о числе каталогов и платформ посчитаны по открытому реестру каталогов данных dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. В нём описаны 43 614 каталогов и 790 платформ. Основные запросы, которыми получены цифры книги, собраны в приложении о реестре: на этом выпуске они дают те же числа. Реестр обновляется каждые несколько дней, на более новых выпусках числа будут другими.

Реестр изначально создавался для поисковика по наборам данных Dateno, и Dateno до сих пор индексирует порталы по их описаниям в реестре: из записи он узнаёт платформу каталога и адреса API, а по ним — каким сборщиком забирать наборы. На 16 октября 2026 года в Dateno 22 миллиона наборов из 5 236 источников. Почему одни каталоги собрать легко, а до других сборщик не доходит, разобрано в главах «Поисковики и агрегаторы» и «Находимость».

Версии программ и возможности платформ проверены по первоисточникам в октябре 2026 года.

Об издании

Автор — Иван Бегтин. **Редакция** — октябрь 2026 года.

Поддержка. Книга и база знаний создаются при поддержке проекта Dateno, поисковой системы по наборам данных. Автор — основатель и технический директор Dateno; как это сказывается на источниках книги, объяснено в главе «Об авторе».

Лицензия. Текст распространяется по лицензии Creative Commons «Атрибуция» 4.0 Международная (CC BY 4.0). Его можно копировать, переводить и использовать в своих материалах, в том числе коммерческих, если указать автора и дать ссылку на книгу.

Как цитировать.

> Бегтин И. Каталоги данных: как устроены порталы открытых данных, геопорталы и репозитории — и как выбрать свой. — Редакция октября 2026 г. — URL: https://datacatalogs.begtin.tech/ (дата обращения: ДД.ММ.ГГГГ).

ЧАСТЬ I.

Что такое каталог данных

Определения, границы понятия, пользователи каталогов, устройство каталога и общая картина по данным реестра.

1. Что такое каталог данных
2. Зачем создают каталоги и кто ими пользуется
3. Анатомия каталога
4. Мир каталогов в цифрах

Что такое каталог данных

Каталог данных — это сервис, который описывает наборы данных и помогает их найти, понять и получить. Описание называют метаданными: название, кто опубликовал, лицензия, период, формат, адрес файла или сервиса.

Сам файл может лежать в каталоге или по ссылке в другом месте. Это главное техническое различие, к нему книга возвращается в главе о хранилище. Без описания ссылка на файл — ещё не каталог.

Портал, репозиторий, агрегатор

Слова рядом часто путают.

Портал — каталог плюс всё, что вокруг него: новости, примеры использования, обсуждения, форма обратной связи. data.gouv.fr — портал: у набора есть страница с примерами использования и обсуждением. Страница со списком из десяти CSV — ещё нет, хотя данные она тоже публикует. Когда такого списка достаточно, разобрано в части «Публикация без портала».

Репозиторий — каталог, который хранит файлы у себя и обычно выдаёт постоянный идентификатор, чтобы на данные сослались в статье. Типичный пример — Zenodo.

Геопортал — каталог слоёв и картографических сервисов, часто вместе с картой в браузере.

Агрегатор или поисковик данных — каталог каталогов. Он не публикует свои наборы, а собирает описания чужих. Google Dataset Search и Dateno устроены так.

Реестр метаданных описывает не наборы, а структуры: словари, схемы, другие каталоги. Это другая полка, хотя слово «каталог» в названии встречается.

Хранилище и озеро данных — где байты лежат внутри организации. Каталог отвечает на другой вопрос: как эти данные найти снаружи.

Три большие семьи

Если отвлечься от деталей, каталоги данных делятся на три семьи. Такое деление предлагалось ещё в 2020 году, и оно до сих пор помогает быстро понять, с чем имеешь дело.

	Порталы открытых данных	Научные репозитории	Корпоративные каталоги
Кто видит	все	все, иногда часть данных по запросу	только сотрудники
Контроль над источниками	почти нет: публикуют ведомства, каталог принимает что дали	задаёт требования к описанию, но не к самим данным	полный: каталог знает каждую таблицу
Главное в описании	лицензия и условия использования	постоянный идентификатор и ссылка для цитирования	владелец, происхождение, бизнес-словарь
Как связаны с соседями	сбор описаний по DCAT, API CKAN и похожим протоколам	сбор по OAI-PMH, выдача DOI	интеграция с хранилищами и инструментами компании

Книга занимается первыми двумя семьями и их родственниками — геопорталами, каталогами показателей и микроданных, ML-каталогами. Подробная классификация по типам — в части III.

Чем книга не занимается

Корпоративные каталоги вроде Collibra, Alation и DataHub ведут данные внутри компании: кто владелец таблицы, откуда она взялась, кому можно смотреть. С публичным порталом у них общее только слово «каталог». Аудитория, открытость и стандарты другие. Публичный каталог описывает файлы и сервисы, которые можно скачать. Корпоративный описывает таблицы баз данных, конвейеры обработки и отчёты, к которым доступ есть только изнутри. Внедряют его, чтобы найти, где лежат персональные данные, помочь аналитикам найти нужную таблицу и инженерам — понять, откуда в ней значения. Общих стандартов описания вроде DCAT такие каталоги почти не используют: им важнее интеграция с инструментами компании. Дальше книга про них не говорит.

Книга также не ищет конкретный набор «бюджет Москвы за 2024 год». Для этого открывают сам портал или поисковик данных. Здесь речь о том, как такие порталы устроены и как выбрать свой.

Короткая история

- 2007 год: восемь принципов открытых государственных данных и директива ЕС INSPIRE о пространственных данных.
- 2009–2010: национальные порталы data.gov и data.gov.uk, распространение CKAN.
- 2010–2011: рядом с государственными появляются общественные порталы, которые активисты делают, не дожидаясь властей. В России так появился «Хаб открытых данных» hubofdata.ru на CKAN — раньше государственных порталов.
- 2012: re3data, реестр научных репозиториев.
- 2014: словарь DCAT становится рекомендацией W3C. 2016: принципы FAIR о находимости и повторном использовании научных данных.
- 2019: директива ЕС об открытых данных. 2023: регламент о наборах данных высокой ценности, которые государственные органы должны отдавать бесплатно и в машиночитаемом виде.
- 2020-е: каталогов становится так много, что появляются каталоги каталогов и общие поисковики. В 2021 году запущен перечень российских каталогов datacatalogs.ru, в 2023 году — открытый реестр каталогов всего мира dataportals-registry (начат в проекте Common Data Index), в 2024 году на его основе — поисковик по наборам данных Dateno. Рядом растут программы, которые читают каталоги без человека.
- 2026: флагманы первой волны меняются. data.gov уходит с CKAN на собственную разработку, data.gov.uk вместо сбора всего подряд показывает на главной отобранные коллекции, а с лета главная называется National Data Library. Прежний каталог при этом работает.

Типы, которые из этого выросли, разобраны в части III.

Источники

- Директива INSPIRE 2007/2/EC
- DCAT 3
- Принципы FAIR
- re3data
- GDS, What's changing on data.gov.uk and why, 25 марта 2026 года; главная data.gov.uk в веб-архиве 1 апреля 2026 года — ещё без имени National Data Library, 2 июля 2026 года — уже с ним
- И. Бегтин, Telegram: три категории каталогов данных, 14.10.2020; перезапуск hubofdata.ru, созданного в 2010–2011 годах, 14.04.2021; hubofdata.ru на CKAN появился раньше государственных порталов, 12.02.2022; запуск datacatalogs.ru, 03.11.2021; Common Data Index и реестр каталогов, 10.04.2023; запуск Dateno, 13.03.2024
- dataportals-registry — репозиторий создан в марте 2023 года, проверено 12 октября 2026 года; Dateno
- И. Бегтин, Data discovery в корпоративном секторе. Часть 1, июль 2023 — зачем компании внедряют каталоги данных

Зачем создают каталоги и кто ими пользуется

От того, зачем каталог завели и кто им будет пользоваться, зависит почти всё остальное: тип, набор функций и платформа. Каталог «чтобы был» обычно становится витриной, которую не обновляют.

Зачем их создают

- **Закон и отчётность.** Орган власти обязан публиковать открытые данные. Статистическая служба отдаёт показатели. Геопортал входит в инфраструктуру пространственных данных.
- **Прозрачность.** Бюджет, закупки, выборы должны находиться без запроса в канцелярию.
- **Условия гранта или журнала.** Научные данные выкладывают с постоянным идентификатором, чтобы на них сослались.
- **Повторное использование.** Одно ведомство перестаёт выпрашивать таблицу у другого по почте.
- **Продажа и сервисы.** Описание открыто, файл или API — за плату или по договору.
- **Витрина.** Портал запускают к дате или проекту. Этот мотив сам по себе не плох, но без обновления и машинного доступа он даёт антипаттерны.

Если мотива два, каталогов часто тоже два. Статистической службе нужны и показатели, и файлы обследований. Один сайт редко хорошо делает оба дела. Разбор — в примере статистической службы.

Три работы портала

Мотив отвечает на вопрос «зачем завели». Полезно ещё спросить, какую работу портал делает для пользователя. У государственных порталов открытых данных их обычно три, и хороший портал делает хотя бы одну всерьёз.

- **Поисковик и агрегатор.** Портал собирает описания из десятков и сотен систем, где данные на самом деле публикуются, и даёт один поиск. Так устроены национальные порталы крупных стран: data.gov собирает каталоги федеральных ведомств, европейский data.europa.eu — национальные порталы. Российский data.gov.ru тоже задумывался как сборщик перечней с сайтов ведомств — как это получилось, разобрано в антипаттернах.
- **Сообщество.** Портал вовлекает пользователей: обсуждения наборов, примеры использования, запросы на данные. Образец — data.gouv.fr. Публикуют там не только ведомства: у общественного проекта Open Food Facts, открытой базы продуктов питания, есть своя страница организации с наборами.
- **Инфраструктура для справочных данных.** Портал — часть государственной цифровой инфраструктуры и раздаёт данные, на которых работают другие системы и бизнес. Во Франции закон о цифровой республике выделил «данные справочного значения»: на 2026 год это девять наборов, среди них национальная база адресов, реестр предприятий Sirene, официальный географический справочник и кадастровый план. Для них действуют

обязательства по качеству и доступности, а перечень опубликован на data.gouv.fr. В России похожую роль играют отдельные порталы справочников, о них — в главе о каталогах метаданных.

От того, какая работа главная, зависят функции. Поисковику нужен сбор чужих описаний и хороший поиск, сообществу — обсуждения и учётные записи, инфраструктуре — стабильные адреса, массовая выгрузка и API, на которые можно опереться.

Кто пользуется

- Люди: жители, журналисты, исследователи, аналитики, сотрудники других органов.
- Разработчики сервисов и приложений.
- Другие каталоги, которые забирают описания по расписанию.
- Поисковики данных: Google Dataset Search, Dateno, OpenAIRE.
- Программы и ИИ-агенты, которые ищут данные через API. Об этом в главе о трендах.

Человек и программа — разные клиенты. Сайт без выгрузки описаний теряет поисковики и вышестоящие порталы. Каталог только для программ, без понятных страниц, не выполняет задачу прозрачности: житель не обязан уметь читать JSON.

Чего ждут разные люди

Люди из списка выше ждут от каталога разного. Удобно делить их на три основные группы и несколько подгрупп. Один человек может быть сразу в нескольких.

Группа	Чего ждёт
Аналитики	свежих данных по главным темам, прежде всего официальной статистики; привычных форматов — CSV, XLSX; возможности открыть данные в своих инструментах без программирования
Исследователи	данных своей дисциплины, открытых или с понятными правилами доступа; DOI у наборов; форматов и инструментов, принятых в их науке; ясного происхождения данных и методики
Разработчики и инженеры данных	API; массовой выгрузки всего каталога; описания схем и структур; современных форматов — сжатого CSV, Parquet; предсказуемых адресов, чтобы встроить загрузку в свои конвейеры
Журналисты	того же, что аналитики, но с меньшим погружением в технику; главное — чтобы данные вообще были доступны
ML-инженеры	того же, что разработчики, плюс больших объёмов для обучения и связи с площадками вроде Hugging Face
Статистики	содержательных метаданных и специальных стандартов: SDMX для показателей, DDI для обследований
Геоаналитики и геоинформатики	поиска по территории, стандартных сервисов ArcGIS и OGC, выгрузки в распространённых геоформатах

Таблица работает как проверка. Посмотрите на свой каталог глазами каждой группы: может ли она решить свою задачу. Портал, бесполезный для всех семи, не спасут ни дизайн, ни число наборов.

Отдельно стоит бизнес. Ему нужен не портал, а конкретные данные: реестры организаций, закупки, показатели, геоподоснова. Найдя первоисточник один раз, компания подключается к нему напрямую и дальше ждёт стабильного канала и человека, которому можно написать, если что-то сломалось. Для бизнеса важнее, что ведомство вообще публикует нужные данные и отвечает на запросы, чем то, на каком портале они лежат.

Что из этого следует для выбора

Запишите одну главную группу пользователей и одно обязательство. Из них получаются функции. Поисковику нужна разметка schema.org. Вышестоящему порталу — выгрузка в его профиле DCAT. Журналу — DOI и строка цитирования. Жителю — понятное название, лицензия и кнопка скачивания.

Функция без такой записи — кандидат на отказ. Как это оформить, описано в главе о выборе функций.

Источники

- Карта возможностей
- Принципы выбора набора функций
- data.gouv.fr: Service public de la donnée — перечень девяти наборов справочного значения, проверено 12 октября 2026 года; страница организации Open Food Facts
- И. Бегтин, Telegram: три функции порталов открытых данных и их аудитория, 27.03.2023; три группы пользователей данных и их ожидания, 07.12.2025

Анатомия каталога

Портал на CKAN, научный репозиторий и геопортал называют свои части по-разному, но устроены одинаково. В каталоге лежат записи о наборах данных. У набора есть описание, ответственный и один или несколько способов получить сами данные. Вокруг наборов — организации, темы и группы, а внутри программы — база описаний, поисковый индекс и хранилище файлов.

Если различать эти части, можно сравнить любые два каталога, прочитать их программный интерфейс и понять, чего не хватает своему. Глава описывает устройство словами DCAT — словаря W3C для описания каталогов. На нём построено большинство порталов открытых данных, а остальные стандарты на него легко переложить. В конце разобрана одна живая карточка: в ней видны почти все части и типичные поломки между ними.

Общая схема

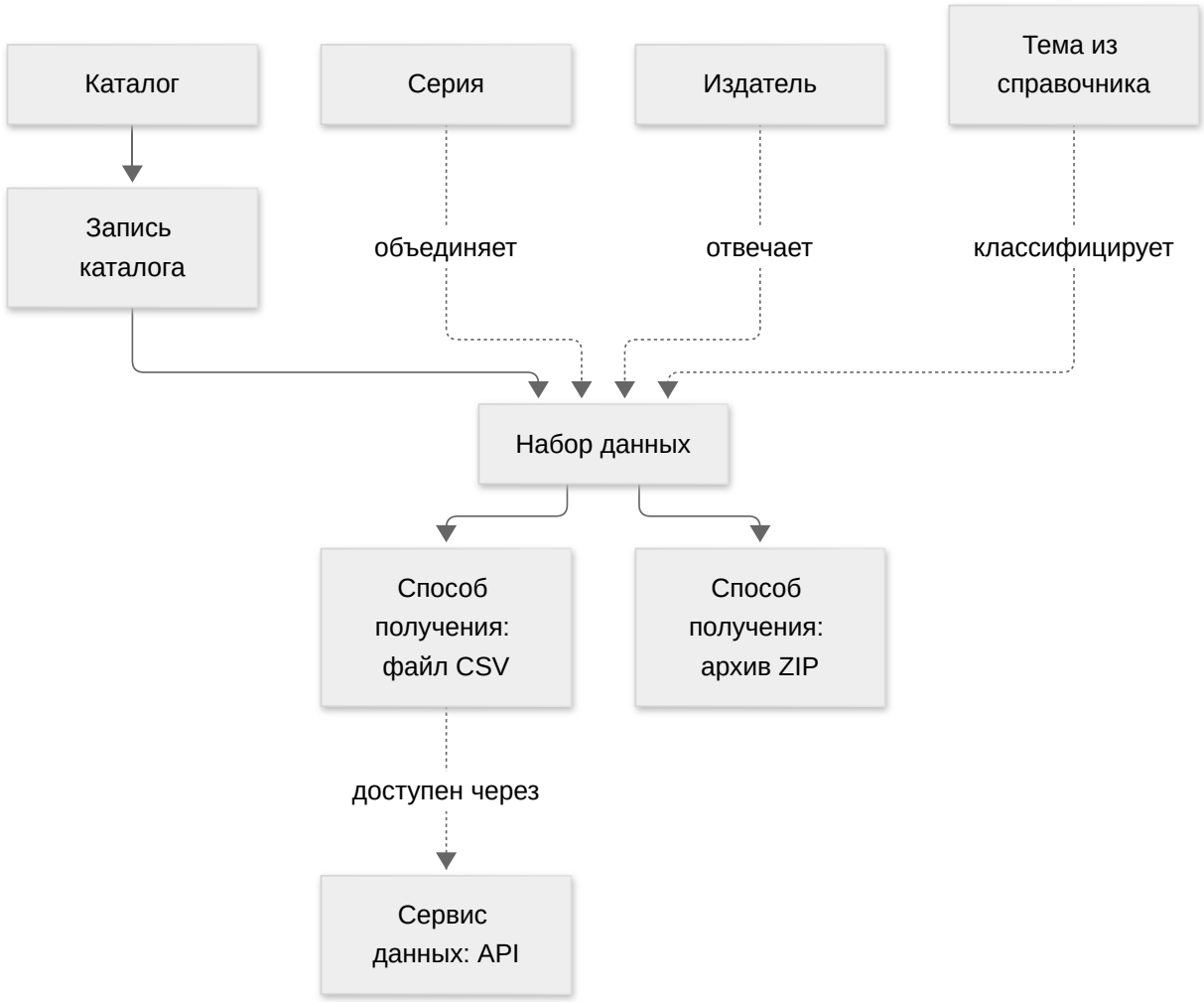


Схема читается сверху вниз: каталог ведёт записи, запись описывает набор, набор отдаётся одним или несколькими способами. Пунктир — связи, которые есть не в каждом каталоге.

Набор данных

Набор данных — то, что пользователь видит как одну карточку и использует как одно целое: бюджет города, слой кадастра, ряд показателя, данные к статье. Что считать набором, решает владелец каталога, и решения сильно различаются: у центрального банка набор — временной ряд, в каталоге спутниковых снимков — коллекция из тысяч файлов. Разбор — в главе о единицах публикации.

Карточка набора отвечает на шесть вопросов. По ним удобно читать любую карточку, даже на незнакомом языке.

Вопрос	Поля	Без чего нельзя
Что это	название, описание, ключевые слова, тема	название и понятное описание
Кто отвечает	издатель, контакт, автор	издатель и способ связаться
О чём и за какое время	период, территория, частота обновления	период — для любых данных, которые меняются
На каких условиях	лицензия, ограничения доступа	лицензия
Как получить	список способов получения	хотя бы один рабочий
Как сослаться	постоянный идентификатор, версия, строка цитирования	постоянный адрес

Какие поля обязательны всем, а какие нужны только отдельным отраслям, разобрано в главе о полноте метаданных.

Тема и тег — разные вещи. Тема берётся из общего справочника: у порталов Европы это список тем Бюро публикаций ЕС, у геопорталов — тематические категории ISO 19115. Тег — свободное слово, которое вписал публикатор. По теме наборы разных порталов складываются в одну полку на общем портале. Теги так не складываются: «транспорт», «Транспорт» и «transport» для программы — три разных слова.

Издатель и организация в программе — тоже не одно и то же. Издатель — тот, кто отвечает за данные по существу и юридически. Организация в программе — учётная запись, внутри которой сотрудники редактируют наборы. Обычно они совпадают. Расходятся, когда наборы за ведомство заводит оператор портала или когда каталог собирает чужие описания: организация тогда — «сборщик», а издатель — исходное ведомство.

Запись о наборе и две пары дат

Запись каталога — сведения не о данных, а о том, как набор попал в каталог: когда его внесли и когда последний раз правили описание. В DCAT для этого есть отдельный класс, и спецификация объясняет его именно на датах. Дата выпуска набора — когда издатель впервые сделал данные доступными. Дата внесения записи — когда о наборе узнал этот каталог.

Большинство программ эти пары дат не различает. Тогда в поле «обновлено» стоит дата последней правки описания, а не данных. Исправили опечатку в названии — и набор пятилетней давности выглядит свежим. Собрали описания с другого портала — и у всех

наборов сегодняшняя дата. Поэтому свежесть данных проверяют по самому файлу и по полю периода, а не по дате на карточке.

Способ получения

Способ получения — конкретный файл, ссылка, API или картографический сервис, через который отдают набор. В DCAT его называют дистрибутивом. У одного набора их может быть несколько: CSV для человека, Parquet для аналитика, API для программы.

У дистрибутива свои поля: формат, размер, контрольная сумма, иногда своя лицензия и своя дата изменения. Две ссылки DCAT различает принципиально:

- **прямая ссылка на файл** (`downloadURL`) — по ней программа одним запросом получает данные;
- **адрес доступа** (`accessURL`) — страница или сервис, через которые данные можно получить, но не обязательно одним запросом.

Если в поле прямой ссылки стоит страница сайта ведомства, человек ещё разберётся, а программа скачает HTML вместо таблицы. Это одна из частых причин, почему поисковик данных показывает набор, но не может определить его формат и размер.

Сервис данных — программный интерфейс, через который данные набора получают по частям: отфильтровать строки, запросить объекты на карте. В DCAT 2 он стал отдельной сущностью: у сервиса есть свой адрес и описание, а дистрибутив ссылается на него. Так удобно описывать картографический сервис, который отдаёт сотню слоёв, или API к таблицам, общий для всего портала.

Серии и версии

Две связи между наборами каталоги путают чаще всего.

Серия объединяет однотипные наборы, которые выходят по очереди: исполнение бюджета за каждый месяц, перепись раз в десять лет. Отдельный класс для серии появился только в DCAT 3 (2024 год), поэтому большинство порталов серий не знает и заводит каждый выпуск отдельной карточкой. На греческом национальном портале data.gov.gr в октябре 2026 года исполнение государственного бюджета — 131 отдельный набор, по одному на месяц с 2014 года, без общей группы. Весь ряд человек находит только поиском по названию. Где граница между серией и раздуванием счётчика — в главе об антипаттернах.

Версия — зафиксированное состояние набора, на которое можно сослаться. Научным репозиториям версии нужны обязательно: статья ссылается на данные, которые были у авторов, а не на последнюю выгрузку. В InvenioRDM, на котором работает Zenodo, для этого три уровня. Родительская запись объединяет все версии и хранит общее для них: права доступа, принадлежность к сообществам. У каждой версии своя запись со своими файлами и своим идентификатором. Мелкие правки описания внутри версии сохраняются как ревизии и новой версии не создают. Порталы открытых данных чаще просто заменяют файл, и прошлое состояние пропадает. Подробнее — в главе об идентификаторах и версиях.

Организации, группы и сам каталог

Над наборами каталоги строят ещё один уровень.

- **Организации** отвечают на вопрос «кто отвечает». В CKAN набор обычно принадлежит ровно одной организации, и от неё зависят права на редактирование.
- **Группы, коллекции, сообщества** отвечают на вопрос «что лежит вместе». В CKAN набор может состоять в нескольких группах. В Dataverse коллекции вкладываются друг в друга: университет — факультет — лаборатория. В InvenioRDM сообщество курирует записи: принимает их или отклоняет.
- **Каталог** — тоже сущность со своим описанием: кто его ведёт, каким справочником тем пользуется, на каких условиях можно брать сами описания. Каталог может включать другие каталоги. Так устроены федерации порталов и STAC, где каталоги вкладываются друг в друга, как папки.

Структуру организаций и групп стоит продумать до запуска. Их короткие имена часто входят в адреса страниц. Документация Dataverse прямо предупреждает: если изменить короткое имя коллекции, старые ссылки на неё сломаются.

Одни и те же части под разными именами

Где	Набор	Способ получения	Как группируют	Что важно знать
DCAT	Dataset; серия — DatasetSeries	Distribution; сервис — DataService	каталог, тема из справочника	общий язык для обмена описаниями между порталами
CKAN	dataset, в API — package	resource: загруженный файл или ссылка	организация, группы, теги	слово package осталось от ранних версий и означает то же, что dataset
Dataverse	dataset	файл	вложенные коллекции	коллекция может содержать и наборы, и другие коллекции
InvenioRDM, Zenodo	запись (record)	файл	сообщества	у каждой версии своя запись, общая родительская — на все версии
Геопортал по ISO 19115	запись метаданных о слое или сервисе	онлайн-ресурс в разделе о распространении	ключевые слова, тематическая категория	у одного слоя часто три способа: картинка WMS, объекты WFS, файл
STAC	коллекция, внутри — элементы (items)	ассет: файл снимка или другой файл	каталоги и подкаталоги	по спецификации STAC коллекция соответствует набору в DCAT

Каталоги показателей и микроданных укладываются в ту же схему, но с другим главным объектом: показатель с разрезами вместо файла, обследование с переменными вместо таблицы. О них — главы о каталогах показателей и о микроданных.

Где лежат байты

Каталог может хранить файл сам, только ссылаться на файл у источника или собирать чужие описания и уводить человека на исходный портал. Ссылка дешевле, но чаще ломается. Научной статье нужна копия у репозитория, а не надежда, что ведомственный сайт не переедет.

«Файл у каталога» не значит «на его сервере». Загруженные на data.gov.gr файлы открываются по адресу портала, а отдаются из облачного хранилища Microsoft Azure. Для пользователя разницы нет. Для владельца это строка в бюджете и вопрос, кто отвечает за резервные копии. Варианты хранения и их поддержка в платформах — в главе о хранилище.

Что внутри программы

Снаружи каталог — сайт. Внутри почти у любой серверной платформы одни и те же части:

- **база описаний** — карточки, пользователи, права; обычно это PostgreSQL;
- **поисковый индекс** — отдельная копия описаний, устроенная для быстрого поиска и фильтров. Если индекс разошёлся с базой, набор существует, но не находится. После переноса и обновлений индекс пересобирают;
- **хранилище файлов** — диск сервера или облачное хранилище;
- **хранилище таблиц** — отдельная база, куда загружены строки таблиц, чтобы показывать их в браузере и отдавать по API;
- **фоновые задачи** — сбор чужих каталогов, загрузка таблиц, проверка ссылок, сборка архивов. Их ставят в очередь и выполняют отдельно от сайта;
- **веб-интерфейс и API** — то, что видят люди и программы.

Платформа	База описаний	Поисковый индекс	Файлы	Что ещё
CKAN 2.12	PostgreSQL	Solr 9 или 8	диск сервера, облако — через расширение	Redis; хранилище таблиц DataStore — по желанию, отдельная база PostgreSQL
Dataverse	PostgreSQL	Solr	диск или облачное хранилище	сервер приложений Payara
InvenioRDM	PostgreSQL 15 и новее	OpenSearch 2.12 и новее	сетевой диск, S3 и другие	Redis или memcached для кэша, RabbitMQ или Redis для очереди задач
Статический сайт	нет	нет или небольшой индекс, который скачивается в браузер	файлы рядом со страницами	генератор, который собирает сайт

Каждая часть — отдельная программа, которую обновляют, резервируют и проверяют. Число частей — хорошая первая оценка того, сколько стоит сопровождение. Подробнее — в главе о развёртывании и сопровождении.

Два входа

Человеку нужны поиск, фильтры, карточка и кнопка «Скачать». Программе — то же самое в машинном виде. Тем же машинным входом пользуются другие каталоги, которые забирают описания по расписанию, и поисковики данных. Хороший каталог отдаёт каждую часть обоими способами.

Человек видит	Программа читает	Как это сделано в CKAN
поиск и фильтры	поисковый API	ядро: `/api/3/action/package_search`
карточку набора	запись в JSON или DCAT	ядро: `package_show`; DCAT — расширение ckanext-dcat, адрес карточки с окончанием `.ttl`, `.rdf` или `.jsonld`
кнопку «Скачать»	прямую ссылку на файл	ядро: адрес у ресурса
таблицу с предпросмотром	API к строкам таблицы	ядро, плагин DataStore: `datastore_search`
список всех наборов	файл со всеми описаниями	расширение ckanext-dcat: `/catalog.ttl`
набор в выдаче поисковика	разметку schema.org на странице, карту сайта	разметка — тоже расширение ckanext-dcat

Оба входа должны говорить одно и то же, а расходятся они часто: на странице есть файл, которого нет в выгрузке, или в выгрузке нет лицензии, которая видна на странице. Человек видит одно, агрегатор — другое. Какие бывают программные входы и как проверить, что они открыты, — в главе «API и массовая выгрузка».

Разбор живой карточки

Возьмём набор «Метеорологические данные станции Парнита» на data.gov.gr — национальном портале открытых данных Греции на CKAN 2.11 (проверено 12 октября 2026 года). Карточка простая, зато в ней видны почти все части из этой главы.

- **Набор.** Издатель — национальная метеослужба Греции; в программе набор принадлежит организации с тем же названием. Тема — «Окружающая среда» из справочника тем ЕС, тегов восемь, на греческом и английском. Лицензия CC BY 4.0, частота обновления — «непрерывно». Набор отмечен как набор высокой ценности по регламенту ЕС 2023/138, категория «Метеорология».
- **Два способа получения.** Первый — файл CSV на 18 МБ с измерениями каждые пять минут. Он загружен в хранилище таблиц, поэтому строки можно запросить через API. Второй — архив ZIP «Все данные», который портал собирает сам расширением ckanext-downloadall: внутри тот же файл и описание набора `datapackage.json`.
- **Сервис данных.** В выгрузке DCAT у файла CSV указан сервис доступа — API хранилища таблиц со своим адресом. Это та самая связка «дистрибутив — сервис данных».
- **Даты.** В выгрузке DCAT дата выпуска набора — 13 мая 2026 года, дата изменения — 11 октября. Обе взяты из записи каталога: когда карточку создали и когда её последний раз меняли.

Если сравнить части между собой, находится пять расхождений.

1. **Два входа говорят разное.** На странице два способа получения, в выгрузке DCAT — один: архива там нет. Человек увидит ZIP, сборщик европейского портала — нет.
2. **Производный файл отстаёт.** CSV на 11 октября 2026 года содержит измерения по этот день, а архив «Все данные» — только по 14 августа. Причина описана в документации расширения: небольшой архив пересобирается, когда меняют карточку, а не когда обновляется файл. Кто скачал «всё одним архивом», получил данные двухмесячной давности.
3. **Описание и содержимое расходятся.** В карточке период начинается 7 апреля 2026 года, а первая строка файла — 1 апреля.
4. **Словаря полей нет.** В хранилище таблиц все 57 колонок записаны как текст, включая температуру и давление, и ни у одной нет описания. Названия вроде `Temp_Dry_5min` понятны метеорологу, остальным — нет. Пропуски обозначены косыми чертами `/////`, и об этом нигде не сказано.
5. **Для поисковиков ничего нет.** Разметки schema.org на странице нет, карты сайта у портала тоже нет.

На странице ни одно из этих расхождений не видно. Все пять находятся за десять минут, если открыть машинную версию той же карточки.

Как разобрать любой каталог

Возьмите один набор и посмотрите на него тремя способами: страница, запись в API, выгрузка описаний. Для CKAN подставьте адрес портала и имя набора из адреса его страницы:

```
# запись о наборе в JSON: поля, ресурсы, организация
curl "https://portal.example/api/3/action/package_show?id=имя-набора"
# тот же набор в DCAT, если установлено расширение ckanext-dcat
curl "https://portal.example/dataset/имя-набора.ttl"
# первые строки и типы колонок, если файл загружен в DataStore
curl "https://portal.example/api/3/action/datastore_search?resource_id=ID-ресурса&limit=5"
```

У других платформ запись о наборе лежит по похожим адресам: у Dataverse — `/api/datasets/:persistentId/?persistentId=doi:…`, у InvenioRDM и Zenodo — `/api/records/ID-записи`. Каталог STAC сам по себе состоит из файлов JSON, и его можно читать прямо в браузере.

Дальше ответьте на пять вопросов:

1. Сколько способов получения и все ли они открываются?
2. Прямая ссылка ведёт на файл или на страницу?
3. Совпадает ли состав страницы и выгрузки?
4. Какая дата стоит на карточке — данных или записи?
5. Есть ли словарь полей и совпадает ли период в описании с содержимым?

Что из этого следует для своего каталога

- Решите, что у вас набор, а что серия, до того как заводить первые карточки. Переделывать сотни отдельных выпусков в серию потом дорого.
- Проверяйте, различает ли платформа прямую ссылку и страницу доступа, даты данных и даты записи. Если нет, договоритесь, какое поле что означает, и запишите это в правила публикации.
- Продумайте организации и группы заранее: их имена уходят в адреса страниц.
- Проверяйте производные файлы — архивы «всё сразу», предпросмотры, копии в хранилище таблиц — так же, как исходные. Они устаревают незаметно.
- Помните, что каждая внутренняя часть программы — отдельная статья расходов на сопровождение.

Источники

- W3C, DCAT 3, рекомендация от 22 августа 2024 года — каталог, запись каталога, набор, серия, дистрибутив, сервис данных, прямая ссылка и адрес доступа
- CKAN: руководство пользователя — наборы, ресурсы, организации, слово package; установка из исходного кода — PostgreSQL, Solr, Redis, DataStore; проверено 12 октября 2026 года
- Dataverse: управление коллекциями — вложенные коллекции, смена короткого имени ломает ссылки; требования к установке — Payara, PostgreSQL, Solr
- InvenioRDM: модель записей — родительская запись, версии, черновики, ревизии; требования к установке — PostgreSQL, OpenSearch, Redis, RabbitMQ, хранилища
- STAC, обзор спецификации — каталог, коллекция, элемент, ассет; соответствие коллекции набору DCAT
- ckanext-downloadall — когда пересобирается архив «Все данные»
- data.gov.gr, Метеорологические данные станции Парнита: запись в API, выгрузка DCAT; проверено 12 октября 2026 года
- Внутреннее хранилище, API и массовая выгрузка

Мир каталогов в цифрах

Цифры ниже — по выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года). Реестр описывает каталоги, а не наборы внутри них. Запросы, которыми получены числа, лежат в приложении.

Главный вывод: записей типа «геопортал» больше, чем всех остальных типов вместе, — 24 752 против примерно 18,8 тысячи. В это число входят и каталоги, где слой находят по описанию, и серверы, которые отдают карту без такого каталога. Портал открытых данных на CKAN — меньшая часть этого мира.

По типам

Тип каталога	Каталогов	Доля
Геопортал	24 752	57%
Научный репозиторий данных	8 404	19%
Портал открытых данных	5 967	14%
Каталог показателей	3 405	8%
Каталог микроданных	326	менее 1%
Каталог метаданных	172	менее 1%
Поисковик данных	124	менее 1%
Каталог API	115	менее 1%
Список наборов данных	107	менее 1%
ML-каталог	100	менее 1%
Маркетплейс данных	56	менее 1%

Ещё несколько десятков записей не вошли в эти типы: общий исследовательский репозиторий и неразобранное. На выводы они не влияют.

Кому принадлежат

- Местные органы власти — около 40% (17,6 тысячи). Из них 14,4 тысячи — геопорталы.
- Университеты и научные организации — 22%.
- Центральные органы власти — 14%, региональные — 12%.
- Бизнес, некоммерческие и международные организации, сообщества — вместе около 10%.

Длинный хвост — муниципальный геопортал на типовой платформе, а не национальный портал, о котором пишут в обзорах.

На чём работают

- Свои разработки — 16% (7,0 тысячи). Среди геопорталов таких только 4%, среди каталогов показателей — 44%.

- ArcGIS Hub — около 3,5 тысячи геопорталов: это каталоги описаний слоёв. ArcGIS Server — около 4,7 тысячи и GeoServer — около 1,5 тысячи отдают слои и сами каталогом описаний не являются. Вместе такие серверы — около 6,1 тысячи записей, почти весь запас, на который геопорталы обгоняют остальные типы. Подробнее — в главе о геопорталах.
- CKAN — около 1 900 порталов открытых данных. Польская e-mapa — 1,9 тысячи геопорталов.
- Дальше по числу установок идут DSpace, GeoNode, GeoNetwork, OpenDataSoft, Gosweb, WordPress, Drupal, Pure.

Муниципалитет редко выбирает платформу в одиночку. Иногда за него решает государство: российский Gosweb — государственная платформа сайтов органов власти. Чаще решает рынок. e-mapa — продукт польской компании Geo-System, на нём работает больше половины гмин страны. Город в Польше оказывается на e-mapa не потому, что сравнивал её с GeoNetwork, а потому что так поступили соседи и так проще подключиться к национальным реестрам. В большинстве стран геопорталы муниципалитетов делят несколько местных поставщиков ГИС, а не мировые платформы.

Где их больше

Больше всего записей у США (около 8 тысяч), затем Польша, Германия, Италия, Франция, Чехия, Канада, Япония, Испания, Великобритания и Россия. Ещё 1,8 тысячи каталогов не привязаны к одной стране.

Порядок стран говорит и о плотности порталов, и о том, где реестр искал тщательнее. Это не рейтинг открытости.

Число каталогов в стране больше всего зависит от её дохода, потом от населения и от того, насколько страна встроена в мировую науку и экономику. Политический режим влияет меньше, чем можно ожидать: в Нидерландах сотни каталогов, в Мьянме почти ничего, а Китай публикует много, особенно научных данных и через провинциальные порталы. Страны, где своих каталогов почти нет, — беднейшие, малые островные и находящиеся в кризисе. Данные о них всё равно есть, но в каталогах международных организаций, а не национальных. Национальные порталы продолжают появляться: в 2026 году их открыли, например, Бутан, Иордания, Оман, Монако и Лихтенштейн.

Машинный доступ и живые сайты

- Программный интерфейс отмечен у 65% каталогов, не отмечен у 34%.
- 868 каталогов не открываются как действующие, 98 выведены из эксплуатации.
- 613 каталогов помечены как национальные: официальный каталог своего типа в стране, а не любой сайт министерства.
- Пометка «DOI» стоит у 2 тысяч каталогов: 1,4 тысячи научных репозиториев и почти 600 геопорталов. Пометку ставят по платформе: она значит, что программа умеет выдавать DOI, а не что каталог их выдаёт. У геопорталов на деле DOI выдают единицы, см. главу об идентификаторах.

Доля без интерфейса — не мелочь для сборщика федерации и для программы, которой нужен один файл всех описаний. Поисковик при этом читает страницу, если на ней есть разметка schema.org и карта сайта.

Чего числа не показывают

Одна запись — один каталог, какого бы он ни был размера. Hugging Face и сайт небольшого города весят одинаково. Реестр неравномерен по странам и типам: где поиск вёлся дольше, записей больше. Перед цитированием в отчёте сверьте число с более новым выпуском реестра.

Насколько число зависит от усилий поиска, видно по самому реестру. В августе 2026 года в нём было около 14,4 тысячи каталогов, к началу сентября — 29,8 тысячи, к выпуску 1.27.0 от 15 октября 2026 года — больше 43 тысяч. Мир за два месяца не изменился: каталоги начали искать по шаблонам адресов в веб-архиве Common Crawl (адрес с `/rest/services` почти всегда означает сервер ArcGIS, с `/geoserver/` — GeoServer) и по поисковикам интернет-устройств вроде Censys, а описывать их помогали ИИ-агенты. Поэтому рост в основном пришёлся на геопорталы, научные репозитории и региональные каталоги, которых раньше просто не видели. Прежние итоги — 10 тысяч каталогов в 2024 году, 1,7 тысячи в 2023-м — говорят о том, сколько удалось найти, а не о том, сколько существовало.

С числом наборов внутри каталогов ещё сложнее: всё зависит от того, что считать набором. В индексе DataCite, где регистрируют идентификаторы DOI для научных данных, летом 2025 года было около 43 миллионов «наборов». Из них 19,8 миллиона — эксперименты одного японского института термоядерного синтеза, каждому из которых присвоен свой DOI, и ещё около 3,5 миллиона — отдельные наблюдения видов из GBIF, хотя наборов в самом GBIF около 115 тысяч. К октябрю 2026 года «наборов» в DataCite стало 76,4 миллиона, и 42,8 миллиона из них — от того же института. Mendeley Data в 2024 году показывал 2,35 миллиона наборов «из GBIF» при 91 тысяче настоящих. Научный репозиторий Геологической службы США ScienceBase насчитывает 18,7 миллиона записей, но большая часть — нарезка нескольких крупных баз по участкам территории. Одно и то же поле «число наборов» у разных каталогов означает вещи, различающиеся на порядки. Подробнее о единице счёта — в главе о единицах публикации.

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. Запросы — в приложении.
- Geo-System, итоги внедрений технологии e-Gmina в 2025 году — e-мара используют больше половины гмин Польши.
- И. Бегтин, распределение геопорталов по странам, Telegram, 27.09.2026 — в большинстве стран лидируют местные поставщики ГИС.
- И. Бегтин, Telegram, о росте реестра: начало реестра, 1,7 тысячи записей, 10.04.2023; 10 тысяч каталогов, 26.08.2024; поиск каталогов по шаблонам адресов в Common Crawl, 22.05.2025; 29,8 тысячи каталогов, 03.09.2026

- И. Бегтин, Telegram, о счёте наборов: DataCite и институт термоядерного синтеза, 21.07.2025; Mendeley Data и GBIF, 11.01.2024; ScienceBase, 01.11.2024; о связи числа каталогов с доходом страны — пост, 09.04.2024; о новых национальных порталах 2026 года — пост, 16.08.2026

ЧАСТЬ II.

Какие данные публикуют

О чём бывают данные в каталогах, откуда они берутся, какой глубины и в каких единицах и форматах их публикуют.

1. Что публикуют в каталогах
2. Единицы публикации и форматы

Что публикуют в каталогах

Короткий ответ: почти всё, что государство, наука, компании и сообщества знают о мире и записывают. Список школ и каждый платёж из бюджета. Расписание автобуса и место, где этот автобус находится прямо сейчас. Перепись населения и сырой сигнал спутника. Геном бактерии и опись архивного фонда.

Тип каталога из части III подсказывает, где что искать: слои — на геопортале, данные к статье — в научном репозитории. Но содержимое шире любого типа. На городском портале лежат и бюджет, и деревья на улицах, и решения комиссий. В общем научном репозитории — таблицы, записи интервью, 3D-модели находок и обученные нейросети.

Содержимое удобно разбирать по пяти вопросам. От ответов на них зависит, какой каталог нужен.

Вопрос	Примеры ответов
О чём данные	бюджет, транспорт, здоровье, климат, культура
Откуда они взялись	учёт в ведомстве, статистика, датчики, исследование, сообщество
Что это за объект	справочник, реестр, записи о событиях, ряд наблюдений, текст, снимок
Насколько подробно	итог по стране, по району, отдельная запись, сырой сигнал
Как часто меняются	один раз, раз в год, каждый день, непрерывно

О чём: темы

Единого списка тем нет. Есть несколько рубрикаторов, и каждый сделан под свою задачу.

- **Темы DCAT-AP** — 13 рубрик, которыми европейские порталы помечают наборы, чтобы их можно было искать вместе: сельское хозяйство и продовольствие; экономика и финансы; образование, культура и спорт; энергетика; окружающая среда; государство и общественный сектор; здравоохранение; международные вопросы; правосудие и общественная безопасность; регионы и города; население и общество; наука и технологии; транспорт. Словарь ведёт Бюро публикаций ЕС.
- **Тематические категории ISO 19115** — рубрики для геоданных: от рельефа и внутренних вод до кадастра и связи. В версии 2003 года их 19, в ISO 19115-1 добавлены ещё две — внеземное пространство и бедствия.
- **Темы INSPIRE** — 34 темы пространственных данных, которые страны ЕС обязаны описывать и отдавать: адреса, кадастровые участки, гидрография, здания, почвы, распространение видов, зоны природного риска и другие.
- **Категории хартии открытых данных G8** 2013 года — 14 областей «высокой ценности», от реестров компаний и преступности до геномов и расписаний транспорта.
- **Категории ODIN** — 22 области официальной статистики, по которым проект Open Data Watch оценивает статистические службы стран: население, образование, здоровье, цены, торговля, энергия, загрязнение и другие.

Рубрикатор нужен, чтобы человек мог сузить поиск, а агрегатор — сложить наборы разных стран на одну полку. Свои рубрики завести можно, но их лучше сопоставить с общими. Так делают национальные порталы ЕС: у них своя рубрикация, а наружу уходят темы DCAT-AP. О словарях подробнее — в главе о стандартах метаданных.

Если свести эти рубрикаторы и посмотреть на живые порталы, складывается такая карта тем. Она не исчерпывающая, но показывает ширину.

Власть и право. Структура органов, контакты, должностные лица. Нормативные акты и их проекты. Голосования и стенограммы парламента. Судебные решения. Итоги выборов по участкам. Обращения граждан. Планы и результаты проверок.

Деньги государства. Бюджет — план и исполнение. Закупки и контракты. Гранты и субсидии. Государственное имущество и долг. Зарплаты и декларации чиновников. Международная помощь.

Экономика и компании. Реестры юридических лиц и их владельцев. Лицензии и разрешения. Финансовая отчётность. Банкротства. Цены, внешняя торговля, рынок труда. Патенты и товарные знаки.

Население и социальная сфера. Перепись и демография. Миграция. Доходы, бедность, пособия. Жильё. Школы и вузы, результаты экзаменов. Больницы и аптеки, реестр лекарств, заболеваемость, регистры клинических исследований.

Безопасность. Преступность, ДТП, пожары, чрезвычайные ситуации.

Территория. Адреса, административные границы, кадастр, градостроительные планы, здания, географические названия, рельеф, снимки с воздуха.

Природа и климат. Качество воздуха, воды и почвы. Выбросы и отходы. Охраняемые территории. Наблюдения за видами. Погода, прогнозы, климатические ряды, спутниковые снимки.

Транспорт и инфраструктура. Маршруты и расписания, положение транспорта в реальном времени, парковки, дороги и их ремонт, аэропорты и железные дороги. Энергосети, связь, водоснабжение.

Энергетика и ресурсы. Производство и потребление энергии, недра, лес, рыба.

Наука. Результаты экспериментов и наблюдений, последовательности генов и структуры белков, обзоры неба, опросы, полевые дневники, код и модели.

Культура и наследие. Музейные коллекции, библиотечные каталоги, архивные описи, оцифрованные рукописи и фотографии, объекты наследия, корпуса текстов.

Справочники. Коды стран, регионов, видов деятельности, валют. Сами по себе они мало кому интересны, но без них наборы разных ведомств не сложить друг с другом.

Откуда берутся данные

Происхождение важно не меньше темы. От него зависят полнота, методика, частота обновления и то, что нужно объяснить в описании набора.

Административные данные. Побочный продукт работы ведомства: реестр лицензий появляется, потому что лицензии выдают, а реестр контрактов — потому что их заключают. Это основа порталов открытых данных. Охват полный — в реестре все, кто получил лицензию. Но поля придуманы для учёта, а не для анализа, и меняются вместе с регламентом.

Официальная статистика. Её собирают специально, по методике: переписи, выборочные обследования, отчётность предприятий. Публикуют итоговые показатели, реже обезличенные микроданные. Без методики цифры легко прочитать неправильно.

Наблюдения и измерения. Метеостанции, датчики качества воздуха, гидропосты, счётчики, спутники. Большие объёмы, длинные ряды, часто поток в реальном времени.

Исследования. Эксперимент, моделирование, полевые работы, опрос. Главное здесь — повторяемость: рядом с данными должны лежать их версия, методика и код обработки.

Тексты и документы. Законы, решения судов, стенограммы, протоколы. Документ сам по себе ещё не данные. Данными его делает структура: разметка статей, ссылки на другие акты, поля «дата, суд, стороны».

Сообщества. OpenStreetMap, Wikidata, наблюдения натуралистов-любителей, которые через iNaturalist попадают в GBIF. Охват бывает шире государственного, качество неравномерное, лицензия своя: у OpenStreetMap, например, ODbL.

Компании. Часть сведений они раскрывают по закону: отчётность эмитентов, выбросы. Часть публикуют сами, когда это выгодно: перевозчик выкладывает расписание, чтобы его показывали навигаторы. Остальное — по договору, за плату или никак.

Производные данные. Прогнозы, модели и реанализы — например, климатический реанализ ERA5, который сводит наблюдения и модель в непрерывные ряды по всей планете. Сюда же индексы, рейтинги и обученные модели машинного обучения. Их публикуют вместе с указанием, из каких исходных данных они получены.

Что это за объект

В одной теме лежат объекты разной природы. От природы зависит, как их хранить, обновлять и показывать.

- **Справочники и классификаторы** — коды и названия. Меняются редко, другие наборы на них ссылаются, поэтому важны версия и дата.
- **Реестры объектов** — компании, школы, лицензии, здания. Отвечают на вопрос «что существует сейчас». Меняются каждый день, нужны устойчивые идентификаторы записей и история изменений.
- **Записи о событиях и сделках** — контракты, платежи, ДТП, вызовы скорой. Копятся без конца, их выкладывают порциями по периодам.
- **Наблюдения и временные ряды** — температура, уровень воды, концентрация примесей по часам.
- **Показатели** — итоговые цифры по территории и году с разрезами.

- **Микроданные** — записи об отдельных людях, домохозяйствах, фирмах после обезличивания.
- **Пространственные данные** — векторные объекты (границы, дороги, здания) и растры (снимки, рельеф).
- **Многомерные массивы** — климатические модели и спутниковые продукты: сетка, время и десятки переменных в одном файле.
- **Тексты, изображения, звук, видео, 3D** — корпуса, оцифрованные рукописи, записи интервью, сканы музейных предметов.
- **Последовательности и спектры** — геномы, белки, результаты масс-спектрометрии. Для них есть отдельные мировые архивы.
- **Код, модели и протоколы** — без них научный результат не повторить, поэтому их публикуют рядом с данными.
- **Описания других данных** — словари, схемы, онтологии, каталоги каталогов. О них глава о каталогах метаданных.

В каких файлах и сервисах всё это отдают — в следующей главе.

Глубина: от итога к записи

Один и тот же сюжет можно опубликовать на разной глубине. Вот ступени для трёх тем.

Тема	Итог	Подробнее	Отдельная запись	Глубже всего
Бюджет	закон о бюджете	исполнение по статьям и месяцам	каждый платёж	контракт с этапами и документами
Транспорт	список остановок	маршруты и расписания	где каждый автобус сейчас	обезличенные поездки по проездным
Здоровье	заболеваемость по региону за год	по больнице и месяцу	обезличенные записи о случаях	медицинские карты, только в защищённой среде

В науках о Земле ступени формализованы. NASA делит спутниковые продукты на уровни от 0 до 4. Нулевой — сырой сигнал прибора. Первый — тот же сигнал с привязкой ко времени и месту и с калибровкой. Второй — физические величины, например температура поверхности моря, в точках съёмки. Третий — те же величины на ровной сетке, за день или за месяц. Четвёртый — результат модели, которая объединила несколько измерений. Аналитику обычно нужен третий уровень, тому, кто проверяет алгоритм, — первый.

Чем глубже ступень, тем больше вопросов можно задать данным, но тем выше риски и требования. На глубине появляются персональные данные и коммерческая тайна, объём растёт на порядки, без API и словаря полей не обойтись. Портал, который публикует только верхнюю ступень, выглядит полным, но не отвечает на вопросы, ради которых данные ищут: кому заплатили, где автобус, в какой больнице.

Как часто меняются

- **Один раз.** Итог исследования, архивная коллекция. Нужны неизменная версия и постоянный идентификатор.

- **Периодически.** Годовой отчёт, месячная статистика. Нужен один ряд с одинаковой структурой, а не новый набор каждый год.
- **Постоянно.** Реестр, который правят каждый день. Нужны снимки на дату и выгрузка изменений.
- **Непрерывно.** Транспорт, датчики воздуха, нагрузка энергосистемы. Файлом это не опубликовать: нужен поток или API, а в каталоге — описание этого потока.

Отраслевые стандарты

Когда тему публикуют многие, появляется общий формат. Он описывает не файл вообще, а предмет: что такое контракт, остановка, наблюдение вида. Данные в таком формате сразу понимают программы и агрегаторы: расписания в GTFS подхватывают картографические сервисы, наблюдения в Darwin Core собирает GBIF.

Тема	Стандарт	Что описывает
Закупки и контракты	OCDS	весь путь контракта: план, торги, заключение, исполнение
Владельцы компаний	BODS	кто в итоге владеет компанией и через какие цепочки
Международная помощь	IATI	кто, кому и на что выделяет помощь
Гранты фондов	360Giving	кто, кому, сколько и на что дал
Бюджет	Fiscal Data Package	строки бюджета с классификациями
Общественный транспорт	GTFS и GTFS Realtime, в Европе NeTEx	остановки, маршруты, расписания, положение машин
Городские заявки	Open311 GeoReport v2	сообщения о проблемах в городе и их статус
Социальные услуги	HSDS (Open Referral)	организации, услуги, адреса, часы работы
Законы	Akoma Ntoso	структура акта, поправки, ссылки на другие акты
Отчётность компаний	XBRL	строки отчётности с машиночитаемыми метками
Статистика	SDMX	показатели с разрезами
Обследования	DDI	анкета, переменные, выборка
Биоразнообразие	Darwin Core	наблюдения и образцы: вид, место, дата
Климат, океан, атмосфера	NetCDF с соглашениями CF	переменные на сетке с единицами и координатами
Культурное наследие	модель Europeana (EDM), IIIF	описание предмета, показ изображений высокого разрешения
Астрономия	FITS и стандарты IVOA	снимки, спектры, каталоги объектов неба
Последовательности ДНК	правила INSDC	геномы и прочтения в общих мировых архивах
Машинное обучение	Croissant	файлы, поля и разбиение набора на части

Для каталога отсюда два вывода. Первый: если для темы есть стандарт, публикуйте в нём, а CSV давайте дополнительно. Второй: каталогу не обязательно понимать каждый стандарт изнутри, но он должен уметь сказать, какому стандарту соответствует файл. В DCAT для этого есть поле `dct:conformsTo`. Профильные платформы идут дальше и проверяют файл валидатором. Об этом — в главе о полноте метаданных.

Что обязаны публиковать

Обязательные перечни задают минимум. По ним видно, что государство считает ключевым.

- **ЕС.** Регламент 2023/138 о наборах данных высокой ценности применяется с 9 июня 2024 года. В нём шесть групп: геоданные, наблюдение Земли и окружающая среда, метеорология, статистика, компании и их владельцы, мобильность. Их отдают бесплатно, в машиночитаемом виде, через API и, где уместно, массовой выгрузкой. Ещё раньше директива INSPIRE обязала описывать и отдавать сервисами 34 темы пространственных данных.
- **США.** OPEN Government Data Act — раздел II закона об обоснованной политике 2018 года, подписан в январе 2019 года. Он делает данные федеральных ведомств открытыми по умолчанию и требует вести полную опись своих данных. Описания собирает федеральный каталог data.gov.
- **Россия.** Распоряжение Правительства № 1187-р от 10 июля 2013 года (в редакции 2022 года) содержит три перечня. Федеральные органы публикуют подведомственные организации, планы и результаты проверок, статистику, вакансии, реестры лицензий. Регионы — реестры лицензий на образовательную, медицинскую и фармацевтическую деятельность, лесной и охотхозяйственный реестры. Третий перечень называет сведения из реестров Росприроднадзора и Росгидромета: отходы, объекты, загрязняющие среду, разрешения на выбросы, прогнозы погоды, данные о загрязнении.

Международные ориентиры шире обязательств. Хартия G8 2013 года назвала первыми четыре группы — национальную статистику, карты, выборы и бюджеты, — а следом 14 областей высокой ценности. Индекс ODIN проверяет 22 области официальной статистики не только на наличие, но и на подробность — по годам и территориям.

У перечней есть слабое место. В них легче вписать то, что просто определить: названия органов, списки лицензий. Перечень можно выполнить формально и не опубликовать ни платежей из бюджета, ни расписаний. Поэтому перечень — нижняя граница, а не карта того, что нужно людям.

Чего обычно нет

Сведений об отдельных людях. Медицинских карт, налоговых данных граждан, записей о получателях пособий. Вместо них публикуют агрегаты, обезличенные микроданные по заявке или синтетические данные — искусственные записи с теми же статистическими свойствами. Ещё один путь — дать работать с записями в защищённой среде, не выдавая их на руки. Об этом главы о каталогах микроданных и о доступе и ролях.

Тайн и всего, что касается безопасности. Коммерческой тайны, подробных схем сетей, объектов критической инфраструктуры, части точных геоданных.

Того, за что берут деньги. Кадастр, метеорологию, реестры компаний во многих странах продавали. Регламент ЕС сделал шесть групп высокой ценности бесплатными, но за пределами ЕС платный доступ к таким данным встречается часто.

Сырых и быстрых данных. Первичные записи и ежедневные ряды оставляют в учётной системе и отдают агрегатом.

Данных компаний о клиентах и услугах. Их публикуют редко и обычно агрегатами. Обмен идёт по договорам, всё чаще в пространствах данных — об этом в главе об ИИ и пространствах данных.

Сопровождения. Словаря полей, методики, истории изменений, списка известных ошибок. Без словаря полей файл формально открыт и практически нем.

Неудачных и промежуточных результатов исследований. В репозиторий попадает то, что легло в статью. Отрицательный результат и черновые выгрузки обычно остаются у автора.

Снятого с публикации. Набор может исчезнуть по решению владельца — после смены системы, руководства или политики. Тогда остаются только копии в веб-архивах и у агрегаторов. Это довод хранить сами данные, а не только ссылку на чужой сайт, и не менять адреса без переадресации.

Что видно по реестру

Реестр dataportals-registry описывает каталоги, а не наборы, и содержимое отмечает грубо: какого рода материалы в каталоге. По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) картина такая. Один каталог может попасть в несколько строк.

Содержимое	Каталогов
Наборы данных	42 297
Слои карт	24 655
Показатели	2 876
Публикации	457
Микроданные	151
Документы	120

Слои почти совпадают с числом геопорталов. Показателей меньше, чем каталогов показателей: не у каждой витрины статистики содержимое размечено. Темы наборов — бюджет, транспорт, климат — по реестру каталогов не посчитать. Для этого нужны поисковики по наборам вроде Dateno или каталог data.europa.eu, где у наборов стоят темы DCAT-AP.

Что это значит для каталога

Перед выбором платформы полезно составить опись: для каждой группы данных записать тему, происхождение, вид объекта, глубину, частоту обновления, отраслевой стандарт и ограничения доступа. По описи видно, какие возможности обязательны.

Если публикуете	Что понадобится	Где подробнее
поток в реальном времени	API или поток и его описание в каталоге	интерфейсы доступа
большие массивы, снимки, модели	хранилище вне каталога, выгрузка частями	хранилище
итог исследования	неизменная версия, DOI, экспорт цитирования	идентификаторы, цитирование

Если публикуете	Что понадобится	Где подробнее
реестр, который правят каждый день	снимки на дату, версии, устойчивые идентификаторы записей	идентификаторы и версии
микроданные и записи о людях	доступ по заявке, роли, журнал выдачи	доступ и роли
данные в отраслевом стандарте	отметка о соответствии стандарту, проверка файла	полнота метаданных
слои карт	картографические сервисы, каталог слоёв	геопорталы
таблицы, тексты, изображения	предпросмотр на странице набора	другие возможности
много тем и ведомств	общий рубрикатор тем и сопоставление со своим	стандарты метаданных

Как от этих требований перейти к функциям и платформе — в главе о выборе набора функций.

Источники

- Словарь тем Data theme Бюро публикаций ЕС, на нём основаны темы DCAT-AP.
- Списки кодов ISO 19115, раздел MD_TopicCategoryCode.
- Темы INSPIRE.
- Хартия открытых данных G8 и техническое приложение, 2013.
- Open Data Inventory: категории и методика, выпуск 2024/25.
- Уровни обработки данных NASA EOSDIS.
- Исполнительный регламент (ЕС) 2023/138 о наборах данных высокой ценности.
- Закон США P.L. 115-435, раздел II — OPEN Government Data Act.
- Распоряжение Правительства РФ от 10 июля 2013 года № 1187-р в редакции от 10 ноября 2022 года.
- Отраслевые стандарты: OCDS, BODS, IATI, 360Giving, Fiscal Data Package, GTFS, NeTEx, Open311 GeoReport v2, HSDS, Akoma Ntoso, XBRL, Darwin Core, соглашения CF, IIIF, IVOA, INSDC.
- ERA5 в климатическом хранилище Copernicus.
- Состав содержимого каталогов — выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года.

Единицы публикации и форматы

«Опубликовать данные» значит разное. В одном каталоге это файл. В другом — сервис карты. В третьем — показатель, у которого пользователь выбирает год и регион. Путать эти единицы вредно: из них получается и число «наборов» в отчёте, и требования к платформе.

Что считают одной публикацией

- **Файл или пачка файлов:** CSV, JSON, XML, Parquet, иногда XLSX.
- **API:** программа забирает те же данные запросом, без скачивания всего архива.
- **Картографический сервис:** картинка слоя (WMS, WMTS) или сами объекты (WFS, OGC API Features). Картинка без объектов — просмотр, не выгрузка.
- **Показатель или куб:** таблица с разрезами по времени и территории. Форматы — SDMX, JSON-stat, PC-Axis.
- **Обследование:** описание переменных и файлы записей. Описание обычно в DDI.
- **Набор для обучения модели:** файлы плюс карточка датасета, всё чаще в формате Croissant.

Один сюжет можно выложить десятком файлов «по году и району» или одним набором с API. Для человека десяток файлов иногда понятнее. Для отчёта «у нас 10 000 наборов» это способ раздуть число. Об этом в главе об антипаттернах.

Что считать набором

Слово «набор данных» в разных каталогах означает разное, и при сравнении каталогов это первое, что нужно выяснить.

- **Временной ряд.** У порталов центральных банков, например ЕЦБ и Банка международных расчётов, наборов мало и они огромные, но у каждого ряда — скажем, инфляции в одной стране — есть свой постоянный адрес. Пользователь обычно ищет ряд по своей стране, а не базу целиком. Поэтому поисковики данных показывают ряды отдельными карточками: в Dateno в 2024 году так было устроено около 30% записей.
- **Слой карты.** Чаще всего это не файл, а адрес сервиса ArcGIS или OGC. Его принято считать набором, но честнее называть слоем.
- **Коллекция снимков.** В каталогах STAC одна коллекция — серия снимков территории за разные даты, внутри тысячи и миллионы файлов. В Open Data Cube словом dataset называют как раз один снимок на дату, а то, что другие каталоги называют набором, — product. Лицензия задаётся на product.
- **Запись наблюдения.** Глобальная база биоразнообразия GBIF собирает около ста тысяч наборов, а внутри них — миллиарды наблюдений видов. Если каждому наблюдению выдать идентификатор и назвать его «набором», счётчик вырастет на порядки. Именно так поступают некоторые агрегаторы, о чём в главе о цифрах.

- **Документ и лента.** RSS и ATOM — данные или нет? На европейском портале data.europa.eu в 2024 году было больше 141 тысячи наборов, которые на деле были лентами ATOM геопорталов.

Рабочий критерий — повторное использование. Файл настроек программы машиночитаем, но набором данных не является: его никто не берёт для своей задачи. Коллекция текстов или фотографий не машиночитаема в обычном смысле, но как обучающая выборка — набор. Набором стоит считать то, что пользователь ищет, скачивает и использует как одно целое.

Набор или дата-продукт

Набор данных — то, что опубликовано. Дата-продукт — то, что опубликовано для конкретных пользователей и сопровождается. Различие пришло из корпоративной работы с данными, но всё чаще применяется к открытым: в 2025–2026 годах о нём пишут, например, создатели облачного хранилища открытых данных Source Cooperative. Признаки дата-продукта:

1. Известно, кто им пользуется и для чего: есть основные сценарии использования, даже если продукт бесплатный.
2. Есть ответственный, который понимает и данные, и пользователей, отвечает на вопросы и вносит изменения по обратной связи.
3. Есть документация: схема данных, примеры кода, ограничения.
4. Доступ устроен несколькими способами сразу: API, массовая выгрузка, отдельные срезы. Доступ может быть и регламентированным — по регистрации или в облачной базе.
5. Качество данных описано в документации и проверяется постоянно, а не один раз перед публикацией.
6. Есть версии и график обновления.
7. Дашборд и интерактивная витрина — дополнение к данным, а не их замена.

Под это описание подходят API Всемирного банка, реестр открытых данных Amazon (каждая запись там — документированный массив с примерами использования), российский адресный реестр ФИАС с полными выгрузками и обновлениями. Не подходят большинство наборов на типовых порталах открытых данных и дашборд с кнопкой «скачать в Excel». По оценке И. Бегтина (август 2026 года), на известных порталах открытых данных регулярно обновляется в лучшем случае 5% опубликованного, остальное выложено один раз.

Каталоги устроены под одно или другое. CKAN, DKAN, GeoNode и GeoNetwork считают наборы и прикреплённые к ним файлы. Huwise (бывшая OpenDataSoft) хранит данные таблицами и даёт к каждому набору API и выгрузки, поэтому ближе к каталогу продуктов. Source Cooperative устроен ещё проще: это хранилище файлов, где карточкой продукта служит файл README с описанием схемы, примерами кода и покрытием. Доступ у каждого продукта свой, зато большие массивы, до сотен терабайт, описаны хорошо, потому что их курируют сами авторы.

Для описания дата-продукта есть открытая спецификация ODPS (Open Data Product Specification, версия 4.1): она описывает продукт с несколькими интерфейсами доступа, условиями и ценой. Применяют её пока мало. Практический список, по которому можно проверить свой набор, — в разделе «Набор как продукт» главы о хороших практиках.

Когда файл пригоден для программы

Шкала «пять звёзд» Тима Бернерса-Ли (2010) идёт от файла под любой лицензией к машиночитаемому открытому формату и дальше к ссылкам на общие идентификаторы. Для практики хватает трёх вопросов.

1. Файл открывается без платной программы. CSV, JSON, XML, Parquet, GeoPackage — да. PDF и скан — нет.
2. У таблицы есть строка заголовков и нет объединённых ячеек «для красоты». XLSX с шапкой на три строки программу ломает, хотя формат формально табличный.
3. Колонки названы и объяснены. Схема вроде Frictionless Table Schema или CSV on the Web говорит, что в колонке число, а не текст.

Пятая звезда — связанные данные со ссылками на общие справочники — нужна, когда наборы разных организаций должны склеиваться сами. Для городского бюджета она не первый приоритет. Первые три вопроса — да.

Какой файл выложить

Формат выбирают по трём вопросам: в чём с данными будут работать пользователи, во что можно преобразовать то, что у вас есть, и сколько у вас места на диске и пропускной способности канала. Для большинства случаев ответ укладывается в короткое правило.

- **Плоская таблица** — CSV в кодировке UTF-8, разделитель — запятая, первая строка — заголовки. Такой файл прочитает любая программа. GitHub показывает его таблицей прямо в браузере, если файл не больше 512 КБ, а файл с точкой с запятой покажет просто текстом.
- **Записи со вложенными списками** — у организации несколько адресов, у контракта несколько этапов — JSON Lines: одна запись JSON на строку. Программа читает такой файл строка за строкой. Один большой JSON или XML обычные библиотеки загружают в память целиком, и файл в несколько гигабайт превращается в проблему.
- **Большая таблица для аналитиков** — дополнительно Parquet. Он хранит типы колонок, поэтому программе не нужно угадывать, где число, а где текст, и сжимает данные по колонкам.
- **Отраслевые данные** — в отраслевом формате: GeoJSON или GeoPackage для геоданных, NetCDF для измерений, SDMX для показателей. CSV можно положить рядом, но не вместо.
- **XLSX** — как дополнительная копия для тех, кто будет работать в редакторе таблиц. Импорт CSV с запятой умеет любой такой редактор.

На деле порталы публикуют иначе. В июле 2025 года поисковик Dateno насчитал у наборов с порталов открытых данных 6,7 миллиона файлов и ссылок. CSV — 15%, XLSX, XML, JSON, ZIP — по 7–8%, но ещё 7% — PDF и столько же — ссылки на HTML-страницы, то есть отчёты и чужие сайты вместо данных. Parquet встретился 1652 раза, меньше 0,025%. По опыту того же поисковика, на каждый CSV, который читается без ручной правки, приходятся два с ошибками экспорта: другим разделителем, кодировкой, лишними строками над заголовком.

Разница в объёме заметная. Один и тот же CSV на 3,5 ГБ после перевода в JSON Lines занял 8,7 ГБ, а в Parquet — 0,93 ГБ (пример из практики поисковика Dateno, 2024). JSON Lines дороже по месту, зато сохраняет то, что не ложится в таблицу.

Сжатие

Текстовый файл больше 20–30 МБ стоит сжать. Архиватор выбирают по тому, кто и как будет читать файл.

- **gzip или zstd** — для одного большого файла, который будут читать программы. Многие инструменты открывают такой файл без распаковки: DuckDB, например, сам узнаёт `.csv.gz` и `.csv.zst` по расширению.
- **ZIP** — для пачки файлов, которые скачивают вместе. Zenodo показывает список файлов внутри ZIP прямо на странице записи. Аналитические программы читают ZIP хуже: тот же DuckDB без расширения zipfs его не откроет.
- **RAR и 7z** — лучше не надо. Это популярные у пользователей архиваторы, но мало какая библиотека умеет читать файл изнутри таких архивов потоком, без распаковки на диск.

Что выбрать

Публикуйте ту единицу, которой пользуются. Показатель официальной статистики не заменяют годовым CSV «для галочки», если международный обмен идёт в SDMX: CSV может быть дополнительным файлом. Слой кадастра не заменяют картинкой карты, если человеку нужно скачать границы. Научные данные к статье фиксируют версией файла, а не «всегда последней выгрузкой».

Формат без описания полей и без лицензии остаётся файлом, а не публикацией.

Источники

- Пять звёзд открытых данных
- Frictionless Table Schema
- JSON Lines, Apache Parquet
- Отображение CSV и TSV на GitHub — разделители, ограничение 512 КБ, проверено 12 октября 2026 года
- Чтение CSV в DuckDB — параметр `compression`: gzip и zstd; расширение zipfs для ZIP
- И. Бегтин, Хорошие и плохие практики публикации данных. Метаданные и форматы файлов, ноябрь 2024 — выбор форматов и сжатия, пример с размерами файлов
- И. Бегтин, Telegram, о форматах в индексе Dateno: июль 2025 года, 09.07.2025; Parquet и CSV с ошибками, 20.03.2024
- И. Бегтин, Telegram, о том, что считать набором: повторное использование как критерий, 22.09.2024; временные ряды, слои, ленты ATOM, 25.11.2024; терминология Open Data Cube, 18.05.2023

- И. Бегтин, Telegram, о дата-продуктах: набор и дата-продукт, 05.08.2026; признаки дата-продукта и каталоги наборов и каталоги продуктов, 12.08.2026; Source Cooperative, 25.08.2026; ODPS 4.1, 08.04.2026
- Radiant Earth, Great Data Products, ноябрь 2025; Open Data Product Specification; Source Cooperative; Open Data Cube
- И. Бегтин, Telegram, каталоги наборов и каталоги дата-продуктов, 12.08.2026 — оценка «регулярно обновляется не больше 5% опубликованного»
- Что публикуют в каталогах

ЧАСТЬ III.

Типы каталогов

Как классифицировать каталоги данных и чем различаются основные типы: порталы открытых данных, геопорталы, научные репозитории, каталоги индикаторов и микроданных, ML-каталоги, реестры метаданных и агрегаторы.

1. Как классифицировать каталоги
2. Порталы открытых данных
3. Геопорталы
4. Научные репозитории данных
5. Каталоги показателей
6. Каталоги микроданных
7. ML-каталоги
8. Каталоги метаданных и реестры
9. Поисковики и агрегаторы
10. Каталоги API, маркетплейсы и списки наборов
11. Сравнение типов и гибриды

Как классифицировать каталоги

Каталог данных описывают ответами на пять вопросов: что в нём публикуют, кто его ведёт, где лежат сами данные, кто может их получить и на чём он сделан. Ответ на первый вопрос — тип каталога. От типа зависят стандарт описания и круг подходящих платформ. Остальные четыре ответа объясняют, почему два каталога одного типа всё равно выбирают и сопровождают по-разному.

Классификация нужна не для порядка на полке. По ней находят каталоги, с которыми стоит себя сравнивать, стандарт, которого ждут агрегаторы, класс платформы и типичные сбои. Портал открытых данных города на CKAN и геопортал того же города на ArcGIS Hub оба «про данные муниципалитета». Но стандарты у них разные, платформы разные, и описания они отдают в разные федерации. Если назвать оба сайта одним словом, сравнивать придётся несравнимое.

Пять вопросов

Вопрос	Варианты ответа	Что из ответа следует
Что публикуют	файлы и реестры, слои карт, данные исследований, показатели, микроданные, наборы для машинного обучения, данные на продажу, описания чужих данных	тип каталога, стандарт описания, класс платформы
Кто ведёт и на каком уровне	центральная власть, регион, город, университет или научная организация, бизнес, некоммерческая или международная организация, сообщество	бюджет и команда, обязательства по закону, кому каталог отдаёт описания дальше
Где лежат данные	в самом каталоге, по ссылке у источника, каталог только собирает чужие описания	можно ли выдать DOI, кто отвечает за мёртвые ссылки, нужно ли своё хранилище
Кто может получить данные	все без условий, после регистрации, по заявке, за плату; описание открыто, а данные закрыты	роли и порядок выдачи, нужен ли учёт заявок, что писать в лицензии
На чём сделан	облачный сервис, готовая платформа на своём сервере, картографический сервер или сервер протокола, сайт на CMS, своя разработка	кто обновляет программу, можно ли добавить функцию, сколько стоит уйти

Ответы не зависят друг от друга. Научный репозиторий бывает у университета и у правительства, открытый и с заявками, на своём сервере и в облаке. Поэтому классификацию удобнее записывать строкой, а не одним словом: «геопортал района, слои на собственном картографическом сервере, открытый доступ, облачный ArcGIS Hub». Каждая часть строки отсылает к своей главе книги. Как записать такую строку для своего каталога, показано в конце главы.

Цифры в главе — по выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года), в нём 43 614 каталогов. Реестр — один из источников и сам иногда ошибается в отнесении к типу. Его ограничения разобраны в главе «Мир каталогов в цифрах».

Что публикуют: тип каталога



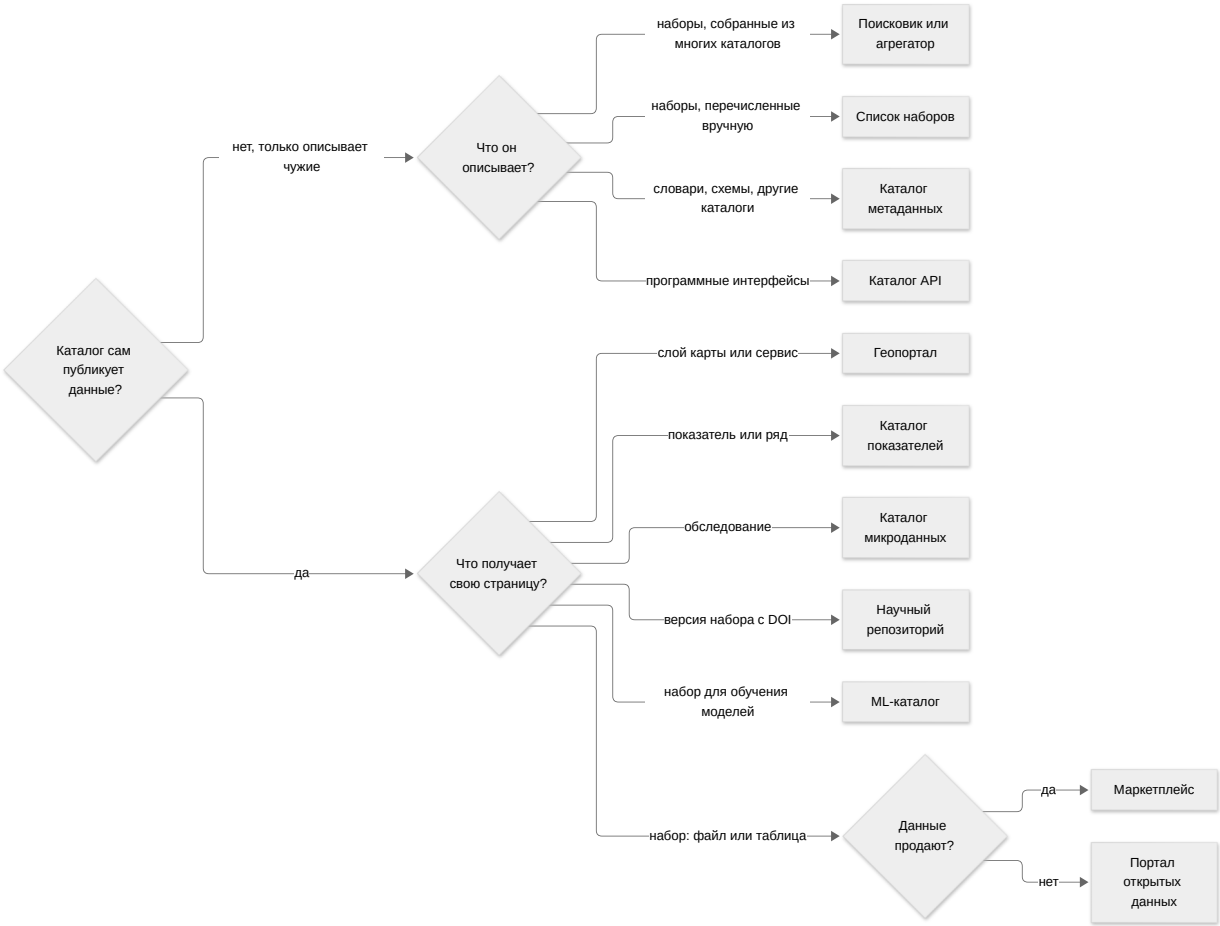
Две ветви различаются главным. Каталоги левой ветви сами публикуют данные. Каталоги правой ведут к чужим данным и отвечают только за описание. Внутри ветви тип определяет главная единица — то, что получает в каталоге свою страницу и свой постоянный адрес.

Тип	Главная единица	Стандарт описания	Пример
Портал открытых данных	набор: файл или таблица	DCAT	data.gouv.fr
Геопортал	слой и картографический сервис	ISO 19115, CSW	Geoportal.de
Научный репозиторий	версия набора с DOI	DataCite	Zenodo, PANGAEA
Каталог показателей	показатель и временной ряд	SDMX	портал данных ЕЦБ
Каталог микроданных	обследование и его файлы	DDI	библиотека микроданных Всемирного банка
ML-каталог	набор для обучения и проверки моделей	Croissant	Hugging Face Datasets
Маркетплейс	товар: набор или подписка по договору	общего нет	AWS Data Exchange
Поисковик или агрегатор	запись о чужом наборе	то, что собирает: DCAT, schema.org, DataCite	data.europa.eu
Каталог метаданных	словарь, схема, описание другого каталога	SKOS, SDMX, ISO/IEC 11179	re3data
Каталог API	программный интерфейс	OpenAPI	api.gov.uk
Список наборов	ссылка на файл	чаще никакого	awesome-public-datasets

Четыре крупных типа — геопорталы, научные репозитории, порталы открытых данных и каталоги показателей — дают 97,5% записей реестра. Остальные семь типов вместе — около тысячи каталогов. Малый тип от этого не становится неважным: каталогов микроданных всего 326, но без них нет доступа к обследованиям, на которых держится социальная статистика.

Как определить тип

Смотрите не на название сайта и не на кнопки главной страницы, а на то, что в каталоге получает собственный адрес. Если у ряда «курс евро к доллару» есть своя страница с графиком и API, а файлов по годам нет, это каталог показателей, даже если сайт называется порталом открытых данных.



Если сайт делает два дела сразу, тип выбирают по тому, что человек видит и делает первым. Карта и каталог слоёв — геопортал, даже если рядом лежит несколько таблиц. Список наборов со встроенной картой — портал открытых данных. Так же стоит выбирать и платформу: не «максимально универсальную», а под главный продукт. Сайты, которые честно делают два дела, и хабы из нескольких каталогов разобраны в главе о сравнении и гибридах.

Кто ведёт и на каком уровне

От владельца зависят деньги, команда и обязательства. Он же часто решает, на какой платформе окажется каталог и куда уйдут его описания: город отдаёт их региону, регион — стране, страна — европейскому или другому общему порталу.

Владельцы в реестре распределены так: местные органы власти — 40%, университеты и научные организации — 22%, центральные органы власти — 14%, региональные — 12%. Бизнесу принадлежит 4% каталогов, некоммерческим организациям — 3%, международным — 3%. У каждого типа свой главный владелец.

Тип	Кто ведёт чаще всего
Геопортал	местные органы власти — 58%, региональные — 14%
Портал открытых данных	местные органы власти — 48%, центральные — 19%, региональные — 14%
Научный репозиторий	университеты и научные организации — 85%

Тип	Кто ведёт чаще всего
Каталог показателей	центральные органы власти — 46%, региональные — 19%, международные организации — 11%
Каталог микроданных	центральные органы власти — 55%, университеты и научные организации — 32%

Что меняется от уровня к уровню:

- **Город.** Наборов мало, людей — один-два. Платформу часто выбирает не город: в России её нередко задаёт государственная платформа сайтов Gosweb, в Польше — рынок, где больше половины гмин работает на e-mapa. Важнее отдать описания наверх, чем завести свою витрину. Разбор — в примере муниципалитета.
- **Регион.** Собирает каталоги городов и отдаёт описания стране. Для него сбор и выгрузка — две обязательные функции.
- **Страна.** Ведёт национальный каталог своего типа и отвечает по закону: директива INSPIRE для геоданных, директива ЕС об открытых данных и регламент о наборах высокой ценности, национальные перечни. В реестре около 600 каталогов отмечены как национальные, хотя отметка расставлена не везде последовательно.
- **Университет или научная организация.** Требования приходят от грантодателей и журналов: DOI, версии, цитирование, сбор описаний по OAI-PMH.
- **Международная организация.** Публикует данные о странах, у которых своих каталогов почти нет, и часто держит несколько каталогов разных типов под одной вывеской.
- **Бизнес.** Открывает данные ради продукта или продаёт их и закрывает доступ, когда это перестаёт окупаться. Так data.world в июле 2026 года закрыл открытое сообщество: подробнее в главе об облачных хранилищах и хабах.
- **Сообщество и некоммерческие организации.** Иногда делают каталог раньше государства, как «Хаб открытых данных» hubofdata.ru в 2010–2011 годах, но держатся на энтузиастах.

Владелец не задаёт тип. 15% каталогов университетов и научных организаций — геопорталы, а не научные репозитории. Владелец не задаёт и охват: международная организация публикует данные о странах, а университет — о территории, где шло исследование.

Где лежат данные

Этот вопрос делит каталоги на три группы.

- **Хранит у себя.** Файл загружен в каталог и живёт, пока жив каталог. Это обычный режим научного репозитория. DOI честнее всего выдавать именно здесь: идентификатор обещает, что по нему найдётся тот же файл и через десять лет, а сдержать обещание может только тот, кто файл хранит.
- **Ссылается на источник.** Каталог хранит описание и адрес, а файл лежит на сайте ведомства или на картографическом сервере. Так устроены многие порталы открытых данных и большинство геопорталов. Риск — мёртвые ссылки: при полной выгрузке data.gov.uk в 2023 году не открывались около 10% ссылок на файлы, к 2026 году — больше четверти.

- **Собирает чужие описания.** Агрегатор и поисковик за данные не отвечают, но отвечают за то, чтобы не показывать мёртвое и не размножать один набор, который есть у города, региона и страны.

Отдельный случай — сервер протокола: THREDDS, ERDDAP, сервер STAC. Данные отдаются по запросу, и каталог с хранилищем здесь совпадают.

Ответ часто зависит не от типа, а от платформы и от решения владельца. CKAN умеет и хранить файл, и ссылаться на чужой. Zenodo рассчитан на то, что файл лежит у него. ArcGIS Hub показывает слои, которые живут на серверах ArcGIS. Варианты и когда какой выбирать — в главе о хранилище данных.

Кто может получить данные

Удобную схему предлагает re3data, реестр научных репозиториев. Он задаёт не один вопрос о доступе, а три:

1. Открыт ли сам каталог: можно ли искать и читать описания.
2. Открыты ли данные. Варианты: открыты, закрыты до определённой даты (эмбарго), выдаются с ограничениями, закрыты совсем.
3. Кто может загружать данные: все, только свои или никто со стороны.

Ограничения он перечисляет отдельно: плата, регистрация, членство в организации. Схема годится не только для науки. Портал может быть открыт для чтения и закрыт для загрузки, и это два разных решения с разными последствиями для ролей и проверки публикаций.

В реестре dataportals-registry 97% каталогов отмечены только как открытые. У 1,4% часть данных открыта, часть выдаётся с ограничениями, у 1,2% доступ только ограниченный. Но по типам разница огромная:

Тип	Доля каталогов с ограниченным доступом хотя бы к части данных
Маркетплейс	91%
Каталог микроданных	49%
Каталог API	14%
Каталог метаданных	10%
Научный репозиторий	6%
Каталог показателей	4%
Портал открытых данных, геопортал	меньше 1%

Ограниченный доступ бывает признаком типа, а не недостатком. Половина каталогов микроданных выдаёт часть файлов по заявке, потому что в обследованиях сведения о людях. Описание при этом остаётся открытым: человек должен знать, что данные существуют, чтобы их запросить. Так устроены UK Data Service и каталог закрытых данных государственного сектора Ирландии, разобранные в главе о каталогах метаданных. У портала открытых данных закрытый доступ, наоборот, — признак беды.

Доступ и лицензия — разные вопросы. Доступ отвечает, можно ли получить данные. Лицензия — что с ними можно делать. Данные бывают общедоступны, но не открыты: смотреть можно, а скачать и использовать повторно нельзя. Примеры — в разделе «Общедоступно, но не открыто» главы о геопорталах, правила — в главе о лицензиях и доступе.

На чём сделан

Класс программы отвечает на вопрос, кто её обновляет, можно ли добавить функцию и сколько стоит уйти. Подробно классы разобраны в главе о классах программного обеспечения. Здесь — как они распределены.

Класс	Доля каталогов	Чаще всего
Облачный сервис	30%	ArcGIS Hub (около 3,5 тысячи), e-mapa (около 1,9 тысячи), OpenDataSoft, Figshare
Готовая платформа на своём сервере или у подрядчика	30%	CKAN, GeoNode, GeoNetwork, DSpace, Dataverse, PxWeb, NADA
Картографический сервер, сервер протокола или отраслевая система	20%	ArcGIS Server (около 4,7 тысячи), GeoServer (около 1,5 тысячи), GBIF IPT, THREDDS, ERDDAP
Своя разработка	16%	—
Сайт на CMS	4%	Gosweb, WordPress, Drupal

По специализированной платформе тип угадывается почти наверняка. CKAN — это 1 906 порталов открытых данных и всего 39 научных репозиториев. DSpace и Dataverse встречаются почти только среди научных репозиториев, GeoNode — только среди геопорталов. Универсальная программа о типе не говорит ничего: WordPress встречается в девяти типах из одиннадцати. Обратное не работает: тип сужает выбор платформы, но не выбирает её. Как от типа и нужных функций перейти к конкретной платформе, разобрано в главе «От функций к платформе».

Другие признаки

Есть признаки, которые не меняют тип, но нужны, когда ищете, с кем себя сравнивать.

Широта темы. re3data делит репозитории на дисциплинарные, институциональные, междисциплинарные, проектные, государственные и прочие. Один репозиторий может попасть сразу в несколько групп. PANGAEA — дисциплинарный, науки о Земле. Harvard Dataverse — институциональный и междисциплинарный одновременно. Zenodo отнесён к прочим. Деление годится и за пределами науки. Отраслевой каталог живёт по стандартам своей области: данные о биоразнообразии описывают в Darwin Core и публикуют через GBIF IPT, океанографические данные отдают серверы ERDDAP. Общий портал страны с ними не конкурирует, а собирает их описания.

Территория. Мир, группа стран, страна, регион, город. Охват не совпадает с владельцем: Всемирный банк публикует обследования десятков стран, GBIF привязывает наблюдение к стране, где оно сделано, а не к стране того, кто его прислал.

Официальный статус. Национальный каталог своего типа — тот, куда по закону или по договорённости собирают данные страны. Его сравнивают с национальными каталогами других стран, а не с порталом соседнего министерства.

Состояние. Работает, заброшен, выведен из эксплуатации, оставлен как архив. В реестре 868 каталогов не открываются как действующие, 98 официально выведены. Сравнивать себя с мёртвым каталогом бессмысленно, но учиться на нём стоит: почему каталоги закрываются и как их спасают, разобрано в главе об архивации.

Ради чего ведут описания. Этот признак объясняет, почему у каталогов одного типа такая разная глубина описания. В 2021 году И. Бегтин выделил пять моделей работы с метаданными:

- научная — делать правильно и неспешно, по стандартам своей дисциплины, чтобы было удобно исследователям этой области;
- архивная — сохранить так, чтобы через 10–20 лет данные можно было понять и использовать снова;
- общественно-государственная — быстро открыть как можно больше; доступность важнее качества, описания минимальные;
- технологическая — связать информационные системы и навязать правила обмена, как делают порталы справочников;
- коммерческая — помочь командам аналитиков обмениваться данными и не терять наработки; это мир корпоративных каталогов.

Научные репозитории и архивы микроданных живут по первым двум моделям, порталы открытых данных — по третьей. Отсюда разница в подробности описаний. Что из этого обязательно для вашего каталога, разобрано в главе о полноте метаданных.

Пограничные случаи: каталог ли это

Сайт можно считать каталогом данных, если выполнены три условия:

1. У набора есть описание: хотя бы название, источник и способ получить данные.
2. Наборы можно найти — поиском, списком или через программный интерфейс.
3. Данные можно получить или хотя бы узнать, как их запросить.

Без первого условия это файловый архив или раздел сайта с вложениями. Без третьего — витрина, на которой данные можно посмотреть, но не забрать; такой случай разобран в антипаттерне «Дашборд вместо данных». Дальше — случаи, где ответ неочевиден.

Репозиторий публикаций, в котором почти нет данных. Университеты ведут на DSpace, ePrints или Pure архивы статей, диссертаций и книг, и иногда туда попадают наборы данных. В архиве открытого доступа СПбГУ в декабре 2024 года было 47 619 записей, из них 17 — наборы данных, то есть 0,03%. DSpace без настройки не фильтрует поиск по типу записи, поэтому, чтобы найти эти 17 наборов, пришлось выгрузить описания всех записей. Такой архив — научный репозиторий, но не репозиторий данных. Его стоит учитывать в перечнях, но не ждать от него того, что ждут от каталога данных. Почему такие архивы трудно собирать, разобрано в главе о научных репозиториях.

Картографический сервер без каталога. ArcGIS Server и GeoServer отдают слои, но сами по себе описаний для поиска не ведут. Реестр учитывает их среди геопорталов: это около 6,1 тысячи записей. Книга считает их сервисами, на которые опирается геокаталог, а не каталогами. Подробнее — в главе о геопорталах.

Портал справочников. Порталы нормативно-справочной информации похожи на порталы открытых данных: много наборов, выгрузка, API. Но у них один владелец, единые условия использования вместо лицензии на каждый набор, и сделаны они для обмена между государственными системами. Решающий вопрос — что человек делает на сайте: скачивает справочник как данные или сверяет по нему коды. Разбор — в разделе «Порталы справочников».

Сайт открытого бюджета. Если таблицы бюджета можно скачать с описанием и датой, это портал открытых данных одной темы. Если на сайте только отчёты в PDF и графики, это сайт с публикациями, а не каталог.

Лента или интерфейс внутри каталога. Ленты ATOM, точки доступа WMS, отдельный API — это способы получить набор, а не новые каталоги и не новые наборы. На data.europa.eu в 2024 году было больше 141 тысячи лент ATOM: так GeoNetwork отдаёт геоданные. Считать ли такой объект данными, решает не его форма, а то, нужен ли он тем, кто работает с данными. Что считать единицей публикации, разобрано в главе о единицах публикации и форматах.

Корпоративный каталог. Collibra, DataHub и похожие системы описывают таблицы внутри компании и снаружи не видны. Это каталоги данных, но не публичные; книга о них не говорит, граница проведена в главе «Что такое каталог данных».

Хаб. Несколько каталогов разных типов под одной вывеской — не новый тип. Каждый каталог хаба классифицируют и оценивают отдельно, см. главу о гибридах.

Типичные ошибки

- **Тип по движку.** «У нас CKAN, значит портал открытых данных». Для специализированных платформ это обычно верно, но на CKAN работают и научные репозитории, а на WordPress — каталоги девяти типов. Тип определяет содержимое, а не программа.
- **Тип по названию.** Слово «портал» или «геопортал» в названии ничего не гарантирует. Сайт статистической службы становится каталогом показателей, когда у каждой таблицы появляется свой адрес и выгрузка, как у Статистики Канады, хотя называется он по-прежнему сайтом службы.
- **Тип по владельцу.** Университет ведёт не только научный репозиторий, но и геопортал, и портал открытых данных. Министерство — не только портал открытых данных, но и каталог показателей.
- **Тип по самой заметной кнопке.** Карта на главной не делает сайт геопорталом, если она показывает уже опубликованные таблицы.
- **Тип по числу наборов.** Миллионы «наборов» часто оказываются нарезкой одной базы или отдельными записями наблюдений. Как это искажает картину, показано в главе «Мир каталогов в цифрах».

- **Один тип на весь хаб.** Показатели и микроданные статистической службы — два каталога с двумя стандартами, даже если домен один.

Как менялись классификации

Деление на типы — рабочий инструмент, и оно менялось вместе с тем, сколько каталогов удавалось найти.

- **2020.** Три большие семьи: порталы открытых данных, научные репозитории и корпоративные каталоги. Эта рамка до сих пор помогает быстро понять, с чем имеешь дело; она разобрана в главе «Что такое каталог данных».
- **2023, февраль.** Каталог каталогов datacatalogs.ru собирал шесть категорий: порталы открытых данных, научные репозитории, базы статистических показателей, геопорталы, сайты открытого бюджета, порталы справочников и классификаторов.
- **2023, май.** В Common Data Index — девять типов: порталы открытых данных, геопорталы, научные репозитории, порталы микроданных и показателей, каталоги данных для машинного обучения, поисковики по данным, маркетплейсы и каталоги API. Порталы справочников туда не вошли: за пределами России похожих порталов почти нет, и тип состоял бы из одних российских сайтов.
- **2026.** В реестре dataportals-registry одиннадцать основных типов. Добавились каталоги метаданных и списки наборов, отдельно учтены общие исследовательские репозитории и неразобранные записи.

Тип заводят тогда, когда каталогов набирается достаточно и их приходится по-разному собирать, описывать и оценивать. Поэтому классификация книги — не окончательная. Если ваш каталог не ложится ни в один тип, опишите его ответами на пять вопросов: этого хватит, чтобы найти соседей и выбрать платформу.

Как записать классификацию своего каталога

Ответьте на вопросы этой главы и запишите ответы в карточку. Её удобно положить рядом с карточками решений из главы «Как обосновать решение».

```
Тип:                  что получает свою страницу
Владелец и уровень:   кто ведёт, город / регион / страна / наука / другое
Где лежат данные:     у себя / по ссылке / только описания
Доступ:               к описаниям, к данным, к загрузке
Класс ПО:             облако / готовая платформа / сервер / CMS / своя разработка
Тема и территория:    общий или отраслевой, какая территория
С кем сравнивать:     три-пять каталогов с теми же ответами
```

Три примера по разобранным в части VI цепочкам выбора:

	Город	Университет	Статистическая служба
Тип	портал открытых данных	научный репозиторий	два каталога: показатели и микроданные
Владелец	местные органы власти	университет	центральный орган власти

	Город	Университет	Статистическая служба
Где лежат данные	файлы в каталоге региона	файлы в самом репозитории	показатели в базе каталога, файлы обследований у себя
Доступ	всё открыто	описания открыты, часть файлов закрыта	показатели открыты, микроданные по заявке
Класс ПО	решает регион: готовая платформа	готовая платформа на своём сервере, рядом с архивом публикаций	две готовые платформы
С кем сравнивать	соседние города того же региона	репозитории других университетов в re3data	службы соседних стран

Подробные решения — в главе «Разобранные примеры». Дальше по ответам выбирают нужные возможности и платформу: с этого начинается маршрутная карта.

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: типы, владельцы, режимы доступа, классы ПО и платформы по типам. Запросы — в приложении.
- Схема метаданных re3data 4.0 и её XML-схема: типы репозиториев, доступ к каталогу, данным и загрузке, виды ограничений. Записи Zenodo, Harvard Dataverse и PANGAEA в re3data, проверено 12 октября 2026 года.
- GDS, What's changing on data.gov.uk and why, 25 марта 2026 года; И. Бегтин, Хорошие и плохие практики публикации данных, ноябрь 2024 — нерабочие ссылки на data.gov.uk.
- И. Бегтин, Telegram: три категории каталогов данных, 14.10.2020; пять моделей работы с метаданными, 10.05.2021; категории datacatalogs.ru, 08.02.2023 (репост канала «Инфокультуры»); порталы НСИ и типы Common Data Index, 22.05.2023; что считать набором данных: ряды, слои, ленты ATOM, 25.11.2024; репозиторий публикаций СПбГУ: 17 наборов из 47 619 записей, 25.12.2024.
- Мир каталогов в цифрах, Сравнение типов и гибриды, Классы программного обеспечения, Карта возможностей

Порталы открытых данных

Портал открытых данных публикует наборы органа власти, города или организации так, чтобы их мог взять и использовать любой. На карточке — таблица, файл, ссылка на API. Карта может быть рядом, но главный продукт — набор, а не слой.

Открытыми данные делают три условия сразу:

- **машиночитаемый вид** — файл или API, который программа читает без ручной перепечатки: CSV, а не скан таблицы в PDF;
- **открытая лицензия** — разрешено копировать, менять и использовать в том числе в коммерческих целях, с одним условием — указать источник;
- **доступ без заявок** — без договора, регистрации и объяснения цели.

Если не выполнено хотя бы одно, это ещё не открытые данные, даже когда раздел сайта так называется. Подробнее о разнице — в главе о геопорталах, в разделе «Общедоступно, но не открыто».

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) порталов открытых данных около 6,0 тысячи, 14% всех каталогов. Действуют около 5,7 тысячи.

Кто создаёт

Почти половину порталов ведут местные органы власти: города, районы, коммуны.

Владелец	Порталов	Доля
Местные органы власти	2 862	48%
Центральные органы власти	1 105	19%
Регионы	818	14%
Университеты и научные организации	465	8%
Компании	268	4%
Общественные организации	264	4%
Международные организации	127	2%

Национальное правительство обычно ведёт один общий портал, рядом бывают порталы отдельных ведомств. Регионы и города — свои или общий региональный. Международные организации и некоммерческие проекты делают то же самое со своими данными. Humanitarian Data Exchange — портал открытых данных Управления ООН по координации гуманитарных вопросов, хотя его владелец не правительство одной страны.

Национального портала может и не быть. В Турции в 2023 году его не было, а открытые данные публиковали города: Стамбул, Измир, Конья. По реестру 1.27.0 общего портала по-прежнему нет, есть ведомственные — Минздрава, избирательной комиссии. В Китае порталы есть почти у каждой провинции, а на национальном уровне с 1 марта 2025 года работает платформа регистрации ресурсов публичных данных. Это реестр, который сводит владельцев

данных с теми, кто хочет их использовать, в том числе за плату, а не портал открытых данных: к августу 2026 года там зарегистрировано больше 300 тысяч ресурсов, продуктов и сервисов.

Число порталов — не мера открытости

По числу порталов в реестре лидируют Россия (572), США (565), Испания (485), Франция (394), Италия (337), Таиланд (311), Индонезия (295), Япония (280) и Китай (227). Список говорит не о том, где открытых данных больше, а о том, как устроена публикация на местах. В России 421 портал из 572 — это раздел «Открытые данные» на типовых сайтах муниципалитетов. В Италии 153 портала — витрины одного облачного сервиса для коммун, в Японии 143 — витрины общего каталога префектур и муниципалитетов, в Таиланде больше 160 — каталоги ведомств и провинций на одной государственной сборке. Сравнивать страны по числу порталов — всё равно что сравнивать по числу сайтов. О том, как такие порталы устроены, — в следующих разделах.

Что публикуют

Бюджеты и закупки, реестры организаций и лицензий, транспорт и расписания, адреса, состояние среды, итоги выборов, статистику. Чем ниже уровень власти, тем короче и проще список. На двух типовых страницах открытых данных районов Калужской области в 2024 году было два и семь наборов: реестры школ, детских садов, библиотек и спортивных организаций, структура администрации. Итальянские коммуны на сервисе Municipium выкладывают десятки наборов, которые сервис сам выгружает из учётных систем: семьи по улицам, рождения, браки, иностранцы по гражданству.

Состав задают обязательные перечни: в ЕС — регламент о наборах данных высокой ценности, в США — OPEN Government Data Act, в России — распоряжение Правительства № 1187-р. Их содержание разобрано в главе «Что публикуют». По перечням видно, что государство считает минимумом, но не то, что нужно пользователям.

На национальных порталах ЕС много геоданных. 12 октября 2026 года общий европейский портал data.europa.eu показывал 1,73 миллиона наборов. Самый частый формат в них — сервис WMS (511 тысяч наборов), за ним WFS (363 тысячи), CSV только на третьем месте (275 тысяч). Крупнейший источник — немецкая инфраструктура пространственных данных GDI-DE: 681 тысяча наборов. Граница с геопорталом здесь размыта: национальный портал открытых данных собирает каталоги слоёв и показывает их рядом с таблицами. Отмечены как наборы высокой ценности около 28 тысяч, 1,6% всего портала.

Как устроены

Внешне порталы похожи: поиск, список наборов, карточка. Устроены они по-разному, и от устройства зависит, какие функции нужны. Чаще всего встречаются пять моделей.

Одна площадка для всех

Ведомства, города, компании и общественные организации публикуют на одном сайте, каждый под своей организацией. Образец — французский data.gouv.fr. 12 октября 2026 года там было 74,4 тысячи наборов, 1265 описанных API, 6,8 тысячи организаций и 81,7 тысячи зарегистрированных пользователей. Портал не только принимает публикации: 507 настроенных сборщиков забирают описания из других каталогов. Он держится на сообществе: 17,7 тысячи обсуждений наборов и около 6 тысяч примеров использования — приложений, статей и исследований, которые пользователи сами привязали к наборам.

Такой модели нужны роли и организации, модерация, обсуждения, учётные записи для внешних издателей и раздел примеров использования.

Сборщик

Портал почти ничего не хранит сам, а собирает описания из каталогов ведомств и регионов и даёт по ним общий поиск. Американский data.gov забирает файлы `data.json`, которые публикуют федеральные ведомства, а также каталоги штатов, городов и университетов: 12 октября 2026 года в нём было 590 тысяч наборов. Европейский data.europa.eu собирает национальные порталы. Индонезийский национальный портал Satu Data в 2022 году собирал региональные порталы на CKAN; в реестре на CKAN работают 196 из 295 индонезийских порталов.

Сборщику нужны сбор по DCAT и другим протоколам, сильный поиск и ссылка на первоисточник в каждой карточке. Свои файлы ему почти не нужны. Как устроены цепочки сбора и где они рвутся — в главе «Федерация и сбор метаданных».

Типовой портал на всех

Государство или поставщик раздаёт одну и ту же сборку десяткам и сотням органов власти. У каждого своя витрина, но программа, поля описания и выгрузка одинаковые. Это ответ на главную беду местного уровня: у маленького города нет людей, чтобы вести свой портал.

- **Таиланд.** Государственный каталог данных GD Catalog — это центральный каталог и каталоги ведомств и провинций на поддоменах `*.gdcatalog.go.th`, все на CKAN с общим тайским расширением. В реестре 164 таких каталога. В октябре 2026 года центральный каталог показывал около 25 тысяч наборов, из них 16,4 тысячи — статистика.
- **Япония.** Сервис BODIK ODCS ведёт исследовательский институт ISIT из Фукуоки. За витринами префектур и муниципалитетов стоит один общий CKAN — data.bodik.jp: в октябре 2026 года в нём было 386 организаций и 18,5 тысячи наборов.
- **Италия.** Municipium — облачная платформа компании Maggioli для сайтов и приложений коммун, портал открытых данных — её часть. Наборы выгружаются из учётных систем коммуны автоматически, описания отдаются в итальянском профиле DCAT-AP_IT национальному порталу dati.gov.it. Общий каталог поставщика на CKAN в октябре 2026 года содержал 6,2 тысячи наборов от 193 организаций.
- **Аргентина.** Andino — сборка CKAN, которую выпустило правительство страны: она упакована в Docker, легко ставится и сразу отдаёт описания в национальном профиле метаданных. В реестре 49 порталов на Andino.

- **Россия.** Gosweb — государственная платформа сайтов органов власти. Раздел «Открытые данные» на ней — страница сайта со списком наборов, где каждый набор лежит тремя CSV-файлами: паспорт, структура и сами данные. Так требовали российские методические рекомендации. API и выгрузки описаний в DCAT у раздела нет, и в реестре ни у одного из 421 такого портала программный интерфейс не отмечен.

Первые четыре примера показывают, что типовой портал работает, когда в нём есть общее хранилище или общая выгрузка описаний и наборы попадают выше без ручной работы. Российский вариант даёт одинаковые страницы, но не общий каталог: найти наборы всех муниципалитетов разом по ним нельзя.

База внутри, выгрузки снаружи

Данные живут в базе или в информационной системе, а портал отдаёт их слепки в нескольких форматах и через API. Наборы обновляются сами, когда меняется база. Такие порталы дороже «файлосборников», но зрелее. Так устроен московский data.mos.ru. Так же — порталы провинций Китая с упором на API реального времени: портал провинции Гуандун в июле 2025 года давал больше 98 тысяч наборов и больше 6,7 тысячи API. Но открытых лицензий там не было, а данные выдавались только после регистрации и проверки личности. По трём условиям из начала главы это общедоступные, а не открытые данные. 12 октября 2026 года из-за рубежа портал не открылся.

Этой модели нужны API, автоматическое обновление наборов из учётных систем и документация полей. Загрузка файлов вручную для неё вторична.

Отобранные коллекции

Портал перестаёт быть складом всего подряд и показывает отобранные и проверенные наборы. В марте 2026 года британский data.gov.uk перестроил главную: на ней теперь коллекции по темам и руководство по работе с данными. Прежний каталог остался отдельным разделом, и служба обещала пока его не менять. Летом 2026 года главная получила имя National Data Library. Причину перестройки назвала сама служба: больше четверти ссылок в каталоге вели на ошибку, многие наборы годами не обновлялись.

Этой модели нужны редакторская работа, проверка ссылок и признаки качества у набора. Это ответ на «свалку», в которую превращается портал, где каждый публикует что хочет и никто не убирает.

Что из этого следует

Модели сочетаются: data.gouv.fr — и площадка, и сборщик, а у типового портала Италии наверху стоит национальный сборщик dati.gov.it. Но главная модель определяет главные функции:

Модель	Без чего не обойтись	Что можно не делать
Одна площадка	роли и организации, модерация, хранение файлов	сбор чужих каталогов на старте
Сборщик	сбор по DCAT, поиск, ссылка на первоисточник	своё хранилище файлов

Модель	Без чего не обойтись	Что можно не делать
Типовой портал	общая схема полей, выгрузка наверх, централизованное обновление	свой дизайн у каждой витрины
База и выгрузки	API, автоматическое обновление, словарь полей	ручная загрузка
Отобранные коллекции	редакторская работа, проверка ссылок, признаки качества	открытая регистрация издателей

Как перейти от этих функций к платформе — в главах «Принципы выбора набора функций» и «От функций к платформе».

Стандарты

Описания отдают в DCAT. В ЕС это профиль DCAT-AP (действующая рекомендация — 3.0.1 от 27 октября 2025 года) и национальные уточнения к нему, например итальянский DCAT-AP_IT. Для наборов высокой ценности есть отдельный профиль DCAT-AP HVD (выпуск 3.0.0, октябрь 2024 года). В США ведомства публикуют файл `data.json` по профилю DCAT-US. Новая версия DCAT-US 3.0 сделана как профиль W3C DCAT 3, поэтому её понимают и зарубежные каталоги. Старую версию 1.1 data.gov собирает на переходный период, новые каталоги рекомендуют сразу делать по 3.0. В России вместо DCAT действует паспорт по методическим рекомендациям версии 4.0; как он устроен — в главе о стандартах метаданных. Поисковикам нужна ещё разметка schema.org на странице набора. Подробнее — в главе об экспорте метаданных.

Программный доступ у каждой платформы свой: Action API у CKAN, API у udata, Explore API у OpenDataSoft, SODA у Socrata. Единого стандарта поиска наборов нет, поэтому вышестоящий портал собирает описания по DCAT, а не через API каждой платформы. Для наборов высокой ценности в ЕС API обязателен, а массовая выгрузка — там, где она уместна. Как проверить их на живом сайте — в главе об API и массовой выгрузке.

Лицензии — CC BY 4.0, ODbL или национальные открытые лицензии вроде британской Open Government Licence и французской Licence Ouverte. Российские порталы публикуют «типовые условия использования общедоступной информации»: по смыслу они разрешают свободное использование, но международным каталогам этот текст незнаком, и программа не распознаёт его как открытую лицензию. Лицензия должна стоять у каждого набора, а не одной строкой в подвале сайта. Подробнее — в главе о лицензиях и доступе.

Платформы

Платформа	Порталов в реестре	Что это	DCAT и API
CKAN	1 906	свободная платформа	API в ядре, DCAT через расширение ckanext-dcat
Своя разработка	1 144	—	API только у 19%, DCAT редко
OpenDataSoft (с 2025 года — Huwise)	454	облачный сервис	в сервисе
Gosweb	421	государственная платформа сайтов органов власти России	список CSV на странице, API нет

Платформа	Порталов в реестре	Что это	DCAT и API
Socrata (Tyler Data & Insights)	245	облачный сервис	в сервисе
DKAN	161	свободная платформа на Drupal	в ядре
Municipium	153	облачная платформа Maggioli для коммун Италии	DCAT-AP_IT для национального портала
BODIK ODCS	143	сервис для муниципалитетов Японии	API общего CKAN
Liferay	117	система управления сайтом, почти все — в Испании	как настроили
EntryScape	99	коммерческая платформа на DCAT, в основном Швеция	в ядре

Реже встречаются Andino (49), JKAN (48), Datasette (44), Data Fair (32, в основном Франция), Piveau (19), udata (15, на нём data.gouv.fr). ArcGIS Hub в реестре записан порталом открытых данных всего 19 раз: гораздо чаще его относят к геопорталам. Это гибрид, о нём в главе о сравнении. Карточки основных платформ — в части «Платформы и расширения».

У национальных порталов и у городских разный выбор. Среди порталов центральных органов власти лидируют CKAN (420) и своя разработка (365), облачных сервисов мало. У местных органов власти CKAN тоже первый (707), но рядом стоят Gosweb (421), своя разработка (307), OpenDataSoft (216), Municipium (153), Socrata (144) и BODIK (127). На местном уровне платформу чаще выбирает не сам город, а государство, регион или поставщик, который уже обслуживает соседей. Так же, как с геопорталами, мировые платформы лидируют в общем счёте, а внутри страны муниципалитеты делят один-два местных поставщика.

Свободные платформы и облачные сервисы почти всегда дают API: в реестре он отмечен у 95% порталов на CKAN и у 98–100% на OpenDataSoft и Socrata. У своих разработок — только у 19%. Это главный довод против самописного портала: интерфейс, который готовая платформа даёт сразу, свою разработку приходится заказывать отдельно, и о нём часто забывают.

Примеры

Национальные порталы

- data.gouv.fr — Франция, своя свободная платформа udata. Образец площадки с сообществом: публикуют ведомства, города, компании и общественные проекты, у наборов есть обсуждения и примеры использования.
- data.gov — США, сборщик. С 2013 года (запуск data.gov 2.0) работал на CKAN, в апреле 2026 года переехал на собственную разработку: сборщик пишет в общую базу, витрина её только читает. Адреса страниц наборов сохранились. Старый Action API CKAN по прежнему адресу отвечает ошибкой, новый интерфейс описан по OpenAPI. Старый каталог временно оставили на catalog-old.data.gov, но 12 октября 2026 года этот адрес уже не открывался. Пример того, что даже флагман CKAN может уйти с платформы, и того, что вместе с платформой меняется API, на который опирались чужие программы.

- data.gov.uk — Великобритания. С марта 2026 года на главной отобранные коллекции, с лета 2026 года она называется National Data Library, прежний каталог работает. Пример перехода от склада к отобранным коллекциям.
- data.gov.au — Австралия, CKAN. С 2019 года поиск и витрину давала австралийская платформа Magda поверх устаревшего CKAN. В июне 2025 года портал перевели целиком на CKAN (на октябрь 2026 года — 2.11.6), а страницы — на Drupal. Больше всего наборов там от национальной геологической службы.
- open.canada.ca — Канада, CKAN, 48 тысяч наборов в октябре 2026 года.

Типовые и местные порталы

- GD Catalog — Таиланд: центральный каталог и каталоги ведомств и провинций на одной сборке CKAN.
- BODIK ODCS — Япония: витрины муниципалитетов над общим CKAN data.bodik.jp.
- Портал открытых данных Стамбула — CKAN. Город публикует данные в стране без национального портала: 286 наборов в 2023 году, 558 в октябре 2026 года.

Международные

- Humanitarian Data Exchange — CKAN 2.11.4 с расширением ckanext-dcat, 28 тысяч наборов в октябре 2026 года. Публикуют агентства ООН, правительства и общественные организации.

Россия

- data.mos.ru — собственная платформа Москвы и пример модели «база внутри, выгрузки снаружи». Один из немногих живых российских порталов, но живёт по инерции: в начале 2022 года новых наборов не было с 2020-го, хотя отдельные реестры обновлялись ежесуточно. В марте 2026 года портал около двух недель был закрыт на «технические работы» без объяснений.
- data.gov.ru — национальный портал. В 2023 году его закрыли «для переноса на новую платформу», летом 2025 года открыли заново без прежних данных. Что было на нём до и после, разобрано в главах о самопроверке каталога и об архивации.

Где ломается

Витрина к дате. Портал запускают к сроку или к отчёту, а потом не обновляют. Снаружи он выглядит живым: главная открывается, поиск работает, но наборы годами одни и те же. Признаки такого портала — в главе о метриках.

Ссылки вместо файлов. Карточка ведёт на файл на сайте ведомства, а через год файл исчезает. На data.gov.uk к 2026 году больше четверти ссылок вели на ошибку — это и стало причиной перестройки. Довод «загружать, а не ссылаться» — в главе о хранилище данных.

Счётчик вместо данных. Если порталу ставят цель «опубликовать N наборов», один ряд режут на десятки файлов по годам и районам, а в зачёт идут таблицы на пять строк. Число наборов растёт, данных не прибавляется. Разбор — в антипаттерне «Дробление ради счётчика».

Лицензия на весь сайт. Условия использования висят в подвале и не указаны у набора. Сборщик забирает карточку без лицензии, и на вышестоящем портале набор уже выглядит закрытым.

Свой формат перечня. Чтобы собрать наборы с тысяч сайтов, придумывают собственную схему вместо DCAT. Так было с российскими перечнями наборов: страна их собирала, а ни один зарубежный агрегатор не читал. Подробнее — в антипаттерне «Свой формат вместо стандарта».

Дверь закрыта. Регистрация, проверка личности, лимиты на API или блокировка запросов из других стран. Каждое из этих условий выбивает портал из поисковиков и агрегаторов. 12 октября 2026 года data.gov.ru, порталы китайских провинций и типовые муниципальные сайты на Gosweb не открывались с зарубежного адреса. Разбор — в антипаттерне «Дверь закрыта для программ».

Своя разработка без API. У самописных порталов программный интерфейс отмечен только у 19%. Национальный или региональный портал без выгрузки DCAT не соберёт нижний уровень и сам не попадёт наверх.

Переезд и закрытие. Около 4% порталов открытых данных в реестре уже не действуют — доля выше, чем у геопорталов и научных репозиториев. Порталы переезжают с платформы на платформу, и при переезде теряются адреса, API и архив. Примеры переездов в обе стороны — в главе о жизненном цикле.

Как проверить портал открытых данных

Пять вопросов, на которые можно ответить за четверть часа, не зная, на чём сделан портал.

1. **Есть ли лицензия у набора?** Откройте три случайные карточки. Лицензия должна быть в каждой, а не только в подвале сайта.
2. **Обновляются ли наборы?** Отсортируйте список по дате изменения. Если свежим записям больше года, портал живёт по инерции.
3. **Можно ли скачать файл напрямую?** Ссылка должна вести на файл или API, а не на страницу ведомства, где файл ещё надо найти.
4. **Отдаёт ли портал описания программам?** Типичные адреса:

```
# CKAN: версия платформы и список расширений
curl -s "https://data.humdata.org/api/3/action/status_show"
# CKAN с ckanext-dcat: каталог в DCAT
curl -sI "https://data.humdata.org/catalog.ttl"
# udata: сводка по порталу
curl -s "https://www.data.gouv.fr/api/1/site/"
# OpenDataSoft: число наборов
curl -s "https://opendata.paris.fr/api/explore/v2.1/catalog/datasets?limit=0"
# США: файл data.json ведомства по DCAT-US
curl -sI "https://www.energy.gov/data.json"
```

5. **Видят ли портал сверху?** Найдите его организацию на национальном портале или в общем европейском data.europa.eu. Если наборов там нет или их заметно меньше, сбор не работает.

Подробный список — в чек-листах, команды для других платформ — в главах об экспорте метаданных и об API.

Когда свой портал не нужен

Маленькому городу или ведомству с парой десятков наборов собственный портал часто вреден: его некому обновлять. Лучше публиковать на региональном или национальном портале, подключиться к типовому порталу, как японские и итальянские муниципалитеты, или выложить наборы статическим каталогом с файлом описаний, который заберёт вышестоящий портал. Разбор такого решения — в примере муниципалитета, способы публикации без портала — в части «Публикация без портала».

Чем отличается от соседей

Сосед	Главное отличие
Геопортал	главный продукт — слой и сервис, а не набор. Если без WMS и WFS сайту нечего показать, это геопортал
Каталог показателей	публикуют файлы и таблицы, а не ряды с разрезами, из которых строят выборку
Научный репозиторий	важнее регулярное обновление, чем неизменная версия с DOI, на которую сослались
Каталог метаданных	портал отдаёт сами данные, а не только описания или справочники
Маркетплейс	данные отдают бесплатно и без договора
Раздел «Открытые данные» на сайте	у портала есть поиск, карточки и выгрузка описаний, у раздела — обычно только список файлов

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: владельцы, страны, платформы порталов открытых данных, в том числе отдельно у центральных и местных органов власти, доля API и недействующих. Запросы — в приложении.
- CKAN: ядро и расширения, карточки платформ
- Объявление о новом каталоге data.gov, 15 апреля 2026 года; описание приложения каталога и его описание API. Число наборов, работа старых адресов наборов и Action API, недоступность catalog-old.data.gov — проверка 12 октября 2026 года.
- DCAT-US 3.0 и переход с версии 1.1; выпуски DCAT-AP, в том числе DCAT-AP HVD 3.0.0
- GDS, What's changing on data.gov.uk and why, 25 марта 2026 года; главная data.gov.uk в веб-архиве 1 апреля 2026 года — ещё без имени National Data Library, 2 июля 2026 года — уже с ним
- data.gouv.fr: сводка по порталу через API; data.europa.eu: поиск по наборам через API — форматы, источники и отметка наборов высокой ценности; open.canada.ca; HDX; портал Стамбула — проверено 12 октября 2026 года
- Типовые порталы: GD Catalog и пример каталога провинции Чайяпхум; BODIK ODCS и общий CKAN data.bodik.jp; Municipium, описание портала коммуны Буссеро на Municipium (автоматическая выгрузка, DCAT-AP_IT, передача в dati.gov.it) и общий каталог Maggioli;

Andino — проверено 12 октября 2026 года

- Раздел «Открытые данные» на Gosweb: снимки веб-архива страниц Боровского района, 14.09.2024, и Козельского округа, 12.08.2024. С зарубежного адреса 12 октября 2026 года сайты не открылись.
- Национальная платформа регистрации ресурсов публичных данных Китая: сообщение о запуске, 01.03.2025; больше 300 тысяч зарегистрированных ресурсов, 25.08.2026
- И. Бегтин, Telegram: три функции порталов, 27.03.2023; data.mos.ru и порталы с базой внутри, 25.10.2021; data.mos.ru: «инерционный сценарий», 17.01.2022; недоступность data.mos.ru в 2026 году, 05.03.2026; data.gouv.fr, 02.04.2025; открытые данные Турции, 07.02.2023; порталы провинций Китая, 24.04.2023; портал провинции Гуандун, 29.07.2025; открытые данные в ЕС, 18.12.2024; о переименовании data.gov.uk, 05.06.2026; о возврате data.gov.au на CKAN, 14.11.2025
- Link Digital, перевод data.gov.au на CKAN; Magda — работала на data.gov.au с 2019 по июнь 2025 года. Версия CKAN — по ответу `/data/api/3/action/status_show`, октябрь 2026 года

Геопорталы

Геопортал описывает пространственные данные — те, что привязаны к месту: слои, карты, снимки, сервисы. У каждого объекта в них есть координаты или граница. Рядом почти всегда есть карта в браузере. Часто портал — видимая часть инфраструктуры пространственных данных страны или региона: общей системы правил, каталогов и сервисов, куда органы власти отдают свои слои.

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) записей этого типа 24,8 тысячи, 57% реестра. Сюда входят и каталоги описаний, и серверы слоёв, и муниципальные карты. ArcGIS Server (4 674) и GeoServer (1 460) отдают карту; сами по себе это ещё не каталог, где слой находят по описанию. Чтобы не путаться, сначала стоит разобрать, из чего геопортал состоит.

Из чего состоит геопортал

Под словом «геопортал» прячутся три разные вещи. Иногда они собраны на одном сайте, иногда живут отдельно.

- **Просмотрщик** — веб-карта. Показывает слои, даёт их включать и выключать, искать адрес, мерить расстояние. Это то, что человек видит первым.
- **Каталог** — карточки слоёв: что изображено, кто сделал, когда, в какой системе координат, с какой точностью, на каких условиях. По карточке слой находят, не зная его адреса.
- **Сервисы** — картографический сервер, который отдаёт слой программам: картинкой для просмотра, объектами для скачивания, готовыми плитками карты.

Часть	Что делает	Стандарты	Примеры программ
Просмотрщик	показывает слои на карте	—	Oskari, TerriaJS, QGIS Web Client, Lizmap, Masterportal
Каталог	описывает слои и ищет по описаниям	ISO 19115, CSW, OGC API Records, DCAT	GeoNetwork, pycsw, ArcGIS Hub
Сервисы	отдаёт сами данные	WMS, WMTS, WFS, OGC API Features, ArcGIS REST, STAC	GeoServer, MapServer, QGIS Server, ArcGIS Server

Есть и платформы «всё в одном»: GeoNode, ArcGIS Online с ArcGIS Hub, российская NextGIS Web.

Открытыми данными слой делает только связка каталога и сервиса со скачиванием. Просмотрщик без них — картинка. Сервер без каталога — слой найдёт только тот, кто знает адрес. Каталог без сервиса — описание без данных. Поэтому и при выборе платформы, и при проверке чужого геопортала первый вопрос один: какие из трёх частей здесь есть.

Кто создаёт

Национальные картографические и кадастровые службы, регионы и особенно города. Геопортал — самый «местный» тип каталога: у города может не быть портала открытых данных, но карта генерального плана, адресов и земельных участков есть почти всегда, потому что с ней каждый день работают сами сотрудники администрации.

Владелец	Геопорталов	Доля
Местные органы власти	14 440	58%
Регионы	3 534	14%
Центральные органы власти	2 752	11%
Университеты и научные организации	1 467	6%
Компании	1 264	5%
Общественные организации	601	2%
Международные организации	392	2%
Прочие	302	1%

Научные и природоохранные организации ведут каталоги наблюдений и снимков, компании — каталоги своих продуктов и открытые слои по заказу государства, международные организации — каталоги по своим программам, например агентства ООН и Европейское агентство по окружающей среде.

Число геопорталов — не мера геоданных

По числу геопорталов в реестре лидируют США (5 050), Польша (2 303), Чехия (1 408), Италия (1 179), Канада (1 027), Германия (964), Франция (899) и Япония (652). Как и с порталами открытых данных, список говорит об устройстве рынка, а не об объёме данных. В США счёт дают ArcGIS Server и ArcGIS Hub: почти 4,5 тысячи записей. В Польше — e-mapa, которой пользуется больше половины гмин. В Чехии — три облачных сервиса для муниципалитетов. Германия при этом только шестая, хотя её общий каталог геоданных — один из крупнейших в мире: вместо тысячи сайтов у неё одна инфраструктура, которая собирает описания всех уровней власти.

Где геопортал — главный каталог

В развивающихся странах геопортал часто главный, а иногда единственный каталог государственных данных. По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) в 47 странах и территориях есть хотя бы один действующий геопортал и нет ни одного действующего портала открытых данных. Больше всего таких геопорталов у Египта (16), Никарагуа (12), Гайаны, Сомали и Гренландии (по 10). Объяснение простое: пользу геоданных легче всего доказать, ими пользуются ведомства, бизнес и международные программы. Поэтому геопортал строят раньше портала открытых данных, часто на деньги донора и на свободном GeoNode.

Бывает и так, что портал открытых данных есть, а данные лежат рядом. В Казахстане, по оценке 2024 года, по-настоящему открытые данные публиковал не портал data.egov.kz, а национальный геопортал map.gov.kz на GeoNode. 12 октября 2026 года его API без входа

отдавал описания 378 наборов.

Что публикуют

Базовые карты и топографию, кадастр и границы участков, адреса, территориальное планирование и зонирование, ортофотопланы — аэро- и космические снимки, приведённые к масштабу карты, — инженерные сети, дороги, охраняемые территории, зоны затопления, геологию, снимки Земли, отсканированные исторические карты.

Единица публикации у геопортала не одна:

- **слой** — одна тема на одной территории, например границы участков города;
- **сервис** — несколько слоёв по одному адресу, например все слои генерального плана;
- **карта или проект** — сохранённый набор слоёв со стилями и подписями;
- **коллекция снимков** — серия снимков одной территории за разные даты.

Один и тот же слой может быть сразу карточкой в каталоге, слоем WMS и WFS на сервере и частью нескольких карт в просмотрщике. Поэтому число «наборов» у двух геопорталов почти никогда не сравнимо: один считает карточки, другой — слои сервера, третий — проекты. Как считать, разобрано в главе «Единицы публикации и форматы».

Как устроены

Внешне геопорталы похожи: карта, список слоёв, поиск. Устроены они по-разному, и от устройства зависит, какие функции нужны. Чаще всего встречаются пять моделей.

Национальная инфраструктура

Ведомства, регионы и города ведут свои каталоги и сервисы, национальный каталог собирает их карточки по CSW и даёт общий поиск, а национальный геопортал показывает слои на одной карте. Образец — Германия. Её общий каталог Geodatenkatalog.de работает на GeoNetwork: 12 октября 2026 года в нём было 814 тысяч записей. Из него же европейский портал data.europa.eu забирает 681 тысячу наборов — больше, чем из любого другого источника. Витрина для людей — Geoportal.de.

Такой модели нужны сбор чужих каталогов, редактор карточек по национальному профилю, проверка карточек на соответствие правилам и наблюдение за тем, отвечают ли сервисы.

Каталог открытых геоданных организации

Город или ведомство выкладывает слои с карточками, скачиванием и лицензией — по сути портал открытых данных, где главный продукт слой. Чаще всего это ArcGIS Hub: сайт поверх облака ArcGIS Online, где каждый слой можно скачать в нескольких форматах. Сайт на Hub сам отдаёт свой каталог в DCAT-US 1.1 и DCAT-AP 2.1.1 и даёт поиск по OGC API Records — например, портал открытых данных Вашингтона opendata.dc.gov; ленты DCAT-US 3.0 и DCAT-AP 3.0.0 включаются отдельно. В реестре такие выгрузки отмечены примерно у 3,2 тысячи сайтов на Hub. Свободная замена — GeoNode: он отдаёт карточки по CSW, но не в DCAT.

Этой модели нужны скачивание в открытых форматах, лицензия у каждого слоя и выгрузка каталога в DCAT, чтобы его забирал портал открытых данных.

Муниципальная карта от поставщика

Облачный сервис местной компании, который подключают десятки и сотни муниципалитетов: участки, генеральный план, адреса, разрешения на строительство. Главный продукт — карта для жителя и чиновника. Слой видно, но скачать его часто нельзя. Кроме ArcGIS Hub и e-mapa, в реестре 132 таких сервиса, на них около 5,3 тысячи геопорталов, и программный интерфейс отмечен лишь у 21% из них. Подробнее о поставщиках — в разделе о платформах ниже.

Такой модели не нужно почти ничего, кроме карты. Но если муниципалитет хочет открытых данных, выгрузку слоёв (WFS или файлы) и лицензию нужно прописывать в договоре с поставщиком: сам сервис их не добавит.

Сервер слоёв без каталога

ArcGIS Server или GeoServer, которые отдают слои программам, но карточек для поиска не ведут. Часто это не отдельный портал, а «задняя часть» других моделей: на сервере держатся карта города и его каталог. Если каталога рядом нет, такой сервер находят только поисковики геосервисов и те, кто знает адрес.

Ему нужен каталог — свой или вышестоящий, — куда слои внесены с описаниями.

Каталог снимков и файлов

Коллекции спутниковых снимков, рельефа, климатических и других пространственных файлов, описанные по STAC. Это не карта города, а склад для программ: файлы лежат в облачном хранилище, каталог отвечает, какие снимки есть на эту территорию за эти даты. Пример — Earth Search компании Element 84, каталог открытых наборов на AWS. В реестре около 140 серверов STAC и около 80 просмотрщиков STAC Browser.

Ему нужны поиск по территории и дате, хранение больших файлов и выдача по прямым ссылкам. Карта для человека вторична.

Что из этого следует

Модели сочетаются: у национальной инфраструктуры внизу сотни серверов слоёв, у сайта на Hub под капотом — ArcGIS Server. Но главная модель определяет главные функции:

Модель	Без чего не обойтись	Что можно не делать
Национальная инфраструктура	сбор по CSW, профиль карточки, проверка карточек, наблюдение за сервисами	своё хранилище данных
Каталог открытых геоданных	скачивание, лицензия у слоя, выгрузка в DCAT	сбор чужих каталогов
Муниципальная карта	просмотрщик, поиск адреса и участка	каталог по ISO, если слои описаны выше
Сервер слоёв	WMS, WFS, регистрация в каталоге	свой просмотрщик
Каталог снимков	STAC, поиск по территории и дате, облачное хранилище	карту для человека

Как перейти от функций к платформе — в главах «Принципы выбора набора функций» и «От функций к платформе».

Стандарты

Описание слоя — ISO 19115 и его XML-вид ISO 19139. Новая редакция стандарта записывается в XML по ISO 19115-3, но каталоги по-прежнему чаще обмениваются карточками ISO 19139. В карточке есть то, чего нет у обычного набора: территория рамкой координат, система координат, масштаб или разрешение, происхождение данных.

Каталог отдают программам сервисом CSW. Ему на смену приходит OGC API Records: часть 1 утверждена как стандарт OGC в апреле 2025 года. Это тот же каталог, но в виде обычного веб-интерфейса с ответами в JSON, который понятен разработчику без знания CSW. Его уже отдают сайты ArcGIS Hub и лёгкий каталог pycsw.

Мост к порталам открытых данных — профиль GeoDCAT-AP. Он описывает, как перевести карточку ISO в DCAT, на котором работают порталы открытых данных. Действующая версия — 3.1.0, рекомендация SEMIC от 16 февраля 2026 года. Именно так data.europa.eu читает национальные геокаталоги. GeoNetwork 4.4 умеет отдавать карточки по CSW сразу в DCAT: каталог Европейского агентства по окружающей среде в октябре 2026 года предлагал GeoDCAT-AP, DCAT-AP и профиль для наборов высокой ценности.

Сами данные отдают сервисами WMS и WMTS для просмотра, WFS и OGC API Features — для скачивания объектов. Картинка карты (WMS) — ещё не данные: из неё нельзя получить границы участков. Скачать объекты позволяют только WFS, OGC API Features, файлы или выгрузка ArcGIS.

ArcGIS REST — отдельный язык той же задачи. Его много не потому, что он стандарт ISO, а потому что на нём сделана большая доля порталов. Программы GIS и многие сборщики читают его наравне со стандартами OGC.

Система координат. У каждого слоя она своя, и указывают её кодом из реестра EPSG. Мировые карты используют WGS 84 (EPSG:4326) и Web Mercator (EPSG:3857), национальные данные — свои системы: S-JTSK в Чехии и Словакии, местные системы координат МСК для каждого региона в российском кадастре. Слой без кода системы координат приходится угадывать.

STAC — протокол для снимков, а за ними и для других наборов файлов с географической привязкой. Это протокол, а не программа. Его отдают отдельные серверы STAC, статические наборы JSON-файлов, но умеют и универсальные платформы: у GeoServer есть модуль сообщества STAC. Поэтому каталогов STAC в мире больше, чем видно по спискам вроде STAC Index. Единица публикации в STAC непривычна для портала: коллекция — это серия снимков одной территории за разные даты, внутри которой тысячи и миллионы файлов. Считать каждый файл набором неверно, а показать коллекцию обычной карточкой с перечнем файлов невозможно. Агрегаторы, которые собирают STAC вместе с порталами открытых данных, решают это каждый по-своему.

INSPIRE

Директива ЕС INSPIRE с 2007 года обязывает органы власти стран ЕС описывать пространственные данные по общим правилам, отдавать каталоги и сервисы просмотра и скачивания. На её основе построены национальные инфраструктуры вроде немецкой, а GeoNetwork стал стандартом для национальных каталогов ЕС.

Сейчас INSPIRE упрощают. В декабре 2025 года Еврокомиссия предложила убрать из директивы технические требования, которые дублируют более поздние законы — директиву об открытых данных и регламент о наборах высокой ценности. Совет ЕС утвердил свою позицию на переговорах в июне 2026 года, Европарламент — 6 октября 2026 года. Парламент хочет сохранить обязательность открытых стандартов ISO и OGC и сеть сервисов. Окончательного текста пока нет.

Одно изменение уже случилось: 1 июля 2026 года геопортал INSPIRE закрыли. Его адрес ведёт на data.europa.eu, где геоданные теперь ищут вместе с остальными открытыми данными ЕС, а словари и коды INSPIRE перенесены в EU Vocabularies Бюро публикаций ЕС. Практический вывод для европейского геопортала: главный путь наверх теперь — описания в DCAT и GeoDCAT-AP, а не только каталог CSW.

Платформы

Только 4% геопорталов самописные. Это самый «платформенный» тип, но и самый пёстрый: в реестре у геопорталов около 350 разных программ.

Платформа	Геопорталов в реестре	Что это	Роль
ArcGIS Server	4 674	коммерческий сервер Esri	сервисы
ArcGIS Hub	3 464	облачный сервис Esri	каталог открытых геоданных
e-mapa	1 927	облачный сервис Geo-System, Польша	муниципальная карта
GeoServer	1 460	свободный сервер	сервисы
Своя разработка	1 043	—	любая; API только у 11%
GeoNode	770	свободная платформа	всё в одном
GeoNetwork	762	свободный каталог	каталог и сборщик
GISPLAN	467	облачный сервис T-MAPY, Чехия и Словакия	муниципальная карта, API нет
Lizmap	346	свободная платформа на QGIS Server, Франция и Италия	просмотрщик и сервисы

GeoServer и ему подобные отдают слои. GeoNetwork и ArcGIS Hub описывают их так, чтобы слой можно было найти. Одно без другого — либо карта без каталога, либо каталог без данных.

Три свободные платформы

У трёх свободных платформ из таблицы разные роли, и при выборе их часто путают.

- **GeoNetwork** — каталог описаний и агрегатор. Он хранит карточки слоёв по ISO 19115, отдаёт их по CSW и сам собирает карточки из других источников: каталогов CSW и OAI-PMH, сервисов WFS и THREDDS, коллекций STAC. В нём есть редактор карточек по правилам INSPIRE, поэтому на GeoNetwork работают многие национальные каталоги ЕС, например эстонский metadata.geoportaal.ee, немецкий Geodatenkatalog.de и каталог Европейского агентства по окружающей среде. Сами данные он обычно не хранит, а ссылается на сервисы. Точнее всего считать его поисковиком по слоям страны или отрасли.
- **GeoNode** ближе всего к обычному каталогу данных. Пользователь загружает файл, и тот становится слоем с карточкой, картой для просмотра, ссылкой на скачивание и API. Внутри работают GeoServer, который отдаёт слой, и pycsw, который отдаёт карточки по CSW. GeoNode выбирают, когда данные публикуют вручную и нужен один сайт «всё в одном».
- **GeoServer** отдаёт слои по WMS, WFS и другим протоколам OGC. Описание слоя у него короткое: название, аннотация, ключевые слова и ссылка на внешнюю карточку. Расширение INSPIRE добавляет в ответ сервиса поля, которых требует директива, но полную карточку ISO GeoServer не ведёт: её держат в GeoNetwork или другом каталоге. Сервер без каталога рядом — это карта без описаний: слой найдёт только тот, кто знает адрес.

Рядом с ними работают витрины поверх чужих каталогов. Финская Oskari, австралийская TerriaJS и QGIS Web Client (QWC2, 159 установок в реестре, больше всего в Финляндии и Германии) показывают на одной карте слои из многих серверов, но своих данных почти не хранят. GeoBlacklight — поиск по коллекциям карт и геоданных университетских библиотек, pycsw — лёгкий каталог описаний. Коротко о каждой — в карточках платформ.

Местные поставщики

e-mapa показывает путь, по которому идёт большая часть муниципальных геопорталов. Это коммерческий продукт польской компании Geo-System: каждая гмина покупает его сама, но поставщик один на большую часть страны, поэтому порталы выглядят одинаково и умеют передавать данные в национальные реестры адресов и планов. Так устроен рынок во многих странах: мировые платформы лидируют в общем счёте, а внутри страны муниципалитеты обслуживают один-два местных поставщика.

Страна	Главные платформы муниципальных геопорталов	Геопорталов в реестре
Польша	e-mapa (Geo-System)	1 927
Чехия	GISPLAN (T-MAPY), GisOnline, Mapotip	424, 287, 261
Словакия	mOBEC (T-MAPY), Geodeticca	177, 159
Словения	iObčina (Kaliopa)	213
Италия	GisMaster (Technical Design)	274
Австрия и Румыния	Map2Web (Schubert & Franzke)	114 и 159
Япония	⊠⊠⊠⊠⊠ (PASCO), SonicWeb	259, 124
Финляндия	Louhi (Sitowise)	117
Австралия	IntraMaps (TechnologyOne), Pozi	111, 80

Страна	Главные платформы муниципальных геопорталов	Геопорталов в реестре
Швейцария	GeoMapFish (свободная)	223
Франция	Lizmap (свободная), mviewer (свободная)	156, 94
Германия	Masterportal (свободная)	85
Китай	Тяньдиту (государственная)	100

В Словении, где 212 общин, у iObčina больше двухсот геопорталов: общины арендуют систему по договору, а данные ведёт команда поставщика. Видна и разница подходов. В Центральной Европе и Японии муниципалитеты арендуют коммерческий облачный сервис, и у большинства таких сервисов программного интерфейса в реестре нет совсем: из 132 сервисов местных поставщиков у 104 он не отмечен ни на одном сайте, в том числе у GISPLAN, GisMaster, iObčina и ⊠⊠⊠⊠⊠. Во Франции, Швейцарии и Германии чаще берут свободную программу, которую развивает сам орган власти или сообщество: Masterportal создала геослужба Гамбурга, mviewer — регион Бретань. Такие программы отдают слои по стандартам OGC, и программный интерфейс отмечен у 84% установок Masterportal и у 99–100% установок Lizmap, GeoMapFish и mviewer.

В России из 286 геопорталов реестра 60 сделаны на своей разработке, 37 — на NextGIS Web, 27 — на ArcGIS Server, 25 — на ActiveMap GIS; GeoServer стоит лишь на 22. Поисковики геосервисов вроде GeoSeer и Spatineo видят только стандартные интерфейсы OGC и ArcGIS и потому охватывают не больше половины таких порталов.

Скрытый каталог ArcGIS

У ArcGIS есть большой каталог, о котором мало кто знает. Любая организация в облаке ArcGIS Online может сделать свои слои, карты, приложения и файлы общедоступными, и всё это попадает в общий поиск ArcGIS. Поиск ArcGIS Hub уже: он ищет по сайтам открытых данных, собранным на Hub. В 2024 году объём общего поиска оценивали в сотни тысяч ресурсов, в 2025 году у Hub было больше 25 тысяч сайтов и около 300 тысяч слоёв, карт и документов. Точнее посчитать нельзя: поиск ArcGIS останавливает счёт результатов на 10 тысячах.

Это полезный, но бедный каталог. У ресурса часто есть только название и теги, без описания, лицензии и даты обновления. Зато в нём находятся слои по странам, где собственных геоданных в сети почти нет: их выкладывают международные проекты, университеты и отдельные специалисты. Многие сайты на адресах `*.maps.arcgis.com` не сообщают, что слои можно скачать, хотя сервер отдаёт их любой программе. Как это проверить — в разделе о неявных интерфейсах.

Примеры

Национальные и международные

- data.europa.eu — с 1 июля 2026 года единая точка поиска геоданных ЕС вместо закрытого геопортала INSPIRE. Собирает национальные геокаталоги, крупнейший из них — немецкий.
- Geoportal.de и Geodatenkatalog.de — витрина и общий каталог инфраструктуры пространственных данных Германии.

- mapy.geoportal.gov.pl — национальный геопортал Польши.
- Каталог Европейского агентства по окружающей среде — GeoNetwork 4.4, около 3,4 тысячи записей, карточки отдаются в ISO и в DCAT.

Города

- opendata.dc.gov — Вашингтон, ArcGIS Hub: скачивание слоёв, каталог в DCAT-US, поиск по OGC API Records.
- Картографический сервис Хельсинки — открытые слои города на GeoServer: в октябре 2026 года по WFS можно было скачать около трёхсот слоёв.

Снимки и файлы

- Earth Search — каталог STAC открытых спутниковых данных на AWS.
- Каталог данных Google Earth Engine — снимки и другие слои в виде каталога STAC, а не карта города.

Россия и Казахстан

- Геопортал Русского географического общества — каталог отсканированных исторических карт, планов и атласов из картографического собрания общества, в котором больше 48 тысяч единиц. Коллекции пополняются: последние добавлены 9 октября 2026 года.
- Геопортал Республики Мордовия — региональный геопортал РГО: тематические слои населённых пунктов, оползней, охраняемых территорий.
- map.gov.kz — национальный геопортал Казахстана на GeoNode, фактически главный источник открытых государственных данных страны.

Где ломается

Карта без скачивания. Слой виден на карте, но выгрузки нет: сервис отдаёт только картинку WMS. Для жителя этого достаточно, для того, кто хочет посчитать или наложить слой на свои данные, — нет.

Карточки для галочки. Поля ISO заполнены прочерками, вместо даты данных стоит дата создания карточки, вместо лицензии — «без ограничений» свободным текстом в поле условий использования. Сборщик такую лицензию не распознаёт, и на вышестоящем портале слой выглядит закрытым.

Каталог и сервис расходятся. Сервис отвечает ошибкой или переехал, а карточка по-прежнему ведёт на старый адрес и этого не показывает. Геопортал Архангельской области потерял слои, когда кончилась лицензия ArcGIS, а каталог продолжал показывать их карточки.

Система координат без кода. Слой отдают в местной системе координат и не указывают её код. Совместить его с другими данными можно только вручную, если вообще можно.

Один поставщик — сотни сайтов. Длинный хвост одинаковых муниципальных порталов никто не обновляет по отдельности: ломается платформа — ломаются сотни сайтов. Уходит поставщик или кончается договор — пропадают сразу все карты.

Свободная платформа без сопровождения. Свой сервер с GeoNode, GeoNetwork или GeoServer кто-то должен обновлять. В реестре не действуют 11% геопорталов на GeoNode и около 7,5% на GeoServer и GeoNetwork, против 1,5–2,3% на ArcGIS Hub и ArcGIS Server. Частая история — геопортал построили по донорской программе, программа кончилась, сервер выключили.

Общедоступно, но не открыто

Слой, который видит каждый, ещё не открытые данные. Открытыми их делают выгрузка в машиночитаемом виде, понятная лицензия и доступ без заявок. Многие государственные геопорталы дают только смотреть. Российские — наглядный пример.

- **Национальная система пространственных данных** Росреестра (nspd.gov.ru). С 23 декабря 2024 года на ней работает публичная кадастровая карта: смотреть можно бесплатно и без входа. Выгрузки слоёв в интерфейсе нет. В январе 2024 года внутри портала нашли GeoServer с более чем 384 слоями: просмотр по WMS был открыт, а ключ к выгрузке по WFS лежал прямо в коде страницы. Вывод того обзора: открытых данных нет, общедоступные есть.
- **ФГИС территориального планирования** Минэкономразвития. Карты генеральных планов и схем территориального планирования органы власти обязаны загружать в векторе (GML) и дополнительно в PDF. Но, как отвечает сама система, векторную модель может получить только обладатель информации, остальным достаются растровые копии. В 2024 году в системе было около 200 тысяч документов, и не меньше чем в четверти из них должны быть векторные геоданные.
- **Фонд пространственных данных Санкт-Петербурга** в 2021 году называл «открытыми» данные, которые выдают бесплатно, но по заявке с паспортными данными и целью использования. Так устроены все государственные фонды пространственных данных в России: данные выдают по заявлению и договору. Бесплатно — ещё не открыто.

Вторая преграда — доступ из-за рубежа. 12 октября 2026 года ни НСПД, ни ФГИС ТП, ни портал петербургского фонда, ни сайт Росреестра не открывались с немецкого адреса. В 2025 году большая часть российских установок GeoNetwork, GeoNode и GeoServer из реестра не отвечала на запросы из-за пределов страны. Такой геопортал не попадает ни в международные поисковики, ни в агрегаторы.

Как проверить геопортал

Пять вопросов, на которые можно ответить за полчаса.

1. **Какие из трёх частей есть?** Посмотрите на адреса страниц и ссылок: `/geonetwork/` — каталог GeoNetwork, `/geoserver/` — сервер GeoServer, `/arcgis/rest/services` — ArcGIS Server, `/api/v2/` рядом с картой — часто GeoNode. Если есть только карта, это просмотрщик.
2. **Можно ли скачать слой?** Ищите кнопку скачивания или сервис WFS. Картинка WMS — не скачивание.
3. **Есть ли у слоя карточка с лицензией и датой данных?** Откройте три случайные карточки. Лицензия должна быть ссылкой на понятный текст, а не «без ограничений».

4. **Отдаёт ли портал описания и данные программам?** Типичные адреса:

```
# GeoNetwork: название каталога и версия программы
curl -s -H "Accept: application/json" "https://sdi.eea.europa.eu/catalogue/srv/api/site"
# Каталог CSW: сколько в нём записей
curl -s "https://gdk.gdi-de.org/gdi-de/srv/eng/csw?
service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=hits"
# GeoServer: какие слои можно скачать по WFS
curl -s "https://kartta.hel.fi/ws/geoserver/avoindata/wfs?
service=WFS&version=2.0.0&request=GetCapabilities"
# ArcGIS Server: папки и сервисы
curl -s "https://maps2.dcgis.dc.gov/dcgis/rest/services?f=json"
# ArcGIS Hub: каталог в DCAT-US и поиск по OGC API Records
curl -s "https://opendata.dc.gov/api/feed/dcat-us/1.1.json"
curl -s "https://opendata.dc.gov/api/search/v1/collections/dataset/items?limit=1"
# GeoNode: число наборов
curl -s "https://map.gov.kz/api/v2/datasets?page_size=1"
# STAC: список коллекций
curl -s "https://earth-search.aws.element84.com/v1/collections"
```

5. **Видят ли портал сверху?** Для Европы найдите его каталог среди источников data.europa.eu или в национальном геокаталоге. Для остальных — поищите его сервисы в GeoSeer. Если портал не нашёлся ни там, ни там, его данные знают только те, кто знает адрес.

Подробный список — в чек-листах, команды для других платформ — в главе об API и массовой выгрузке.

Когда свой геопортал не нужен

Маленькому муниципалитету с десятком слоёв собственный сервер с GeoNode или GeoServer чаще вреден: его некому обновлять, и он закрывается вместе с проектом, на деньги которого построен. Лучше отдать слои в региональную или национальную инфраструктуру, подключиться к сервису, которым уже пользуются соседи, и записать в договор выгрузку и лицензию, или выложить слои файлами GeoPackage и GeoJSON на портал открытых данных с описанием в карточке.

Научному проекту, у которого пространственные данные — результат исследования, нужен не геопортал, а научный репозиторий: файл с DOI переживёт грант, а сервер — нет. Для снимков и других больших файлов достаточно статического каталога STAC в облачном хранилище, без сервера. Разбор решения для города — в примере муниципалитета, способы публикации без портала — в главах о статических сайтах и об облачных хранилищах.

Чем отличается от соседей

Сосед	Главное отличие
Портал открытых данных	главный продукт — набор-файл. Если на сайте главное — карта и каталог слоёв, это геопортал, даже когда рядом лежат таблицы. Гибриды — сайты на ArcGIS Hub и data.europa.eu, где слои и таблицы рядом
Научный репозиторий	хранит неизменную версию данных исследования с DOI. Геопортал показывает актуальное состояние территории и обновляет его

Сосед	Главное отличие
Каталог метаданных	GeoNetwork без своих данных близок к нему, но его карточки ведут к живым сервисам, а не только описывают
Поисковик и агрегатор	GeoSeer и подобные собирают адреса чужих геосервисов и своих карточек не ведут
Просмотрщик карт	показывает слои, но без каталога и скачивания каталогом не является

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: владельцы, страны, платформы геопорталов с долей API и недействующих, классы ПО, интерфейсы, платформы по странам, страны с геопорталом без портала открытых данных, платформы российских геопорталов. Запросы — в приложении.
- Модуль STAC для GeoServer — модуль сообщества, не ядро; STAC Index
- И. Бегтин, Распространение STAC сервисов для геоданных, апрель 2026 — STAC как протокол и трудности с единицей публикации
- Карточки платформ
- Geo-System, порталы на технологии e-mapa — около 2000 гминных и повятовых порталов; итоги внедрений в 2025 году
- Местные поставщики: iObčina (аренда системы общинами по договору), GISPLAN, GisMaster, Map2Web, ⊠⊠⊠⊠⊠, Louhi, IntraMaps; свободные — Lizmap (MPL-2.0), GeoMapFish, Masterportal, mviewer
- И. Бегтин, ПО геопорталов в реестре, Telegram, 17.09.2026, и распределение геопорталов по странам, 27.09.2026
- И. Бегтин, почему геоданные «поглощают» открытые данные, Telegram, 18.04.2023; геопортал Малави и Африка без порталов открытых данных, 17.05.2023; открытые данные Казахстана, 11.09.2024. Число наборов map.gov.kz — по ответу `/api/v2/datasets`, 12 октября 2026 года.
- Документация: сбор карточек в GeoNetwork, редактирование карточек INSPIRE в GeoNetwork, GeoNode, расширение INSPIRE для GeoServer, возможности pycsw
- И. Бегтин, GeoNetwork, GeoNode и GeoServer: роли и число установок, Telegram, 09.06.2025 — оценка ролей платформ и недоступность российских установок из-за рубежа
- Стандарты: OGC, OGC API — Records — Part 1: Core утверждён как стандарт (утверждён 7 апреля 2025 года, текст стандарта); SEMIC, GeoDCAT-AP 3.1.0, рекомендация от 16 февраля 2026 года
- INSPIRE: Еврокомиссия, предложение об изменении директивы 2007/2/EC, COM(2025) 985, 10.12.2025; карточка процедуры 2025/0393(COD) в Европарламенте — позиция Парламента 6 октября 2026 года; позиция Совета ЕС на переговорах, июнь 2026 года; data.europa.eu, INSPIRE datasets move to the European Data Portal, 22.06.2026; Бюро публикаций ЕС, INSPIRE в EU Vocabularies. Перенаправление inspire-geoportal.ec.europa.eu на data.europa.eu — проверка 12 октября 2026 года.

- Проверки 12 октября 2026 года: число записей в Geodatenkatalog.de и в каталоге EEA по CSW, форматы вывода CSW в каталоге EEA (GeoNetwork 4.4.9); наборы GDI-DE на data.europa.eu — поиск по каталогу gdi-de; выгрузки DCAT-US, DCAT-AP и OGC API Records на opendata.dc.gov; слои WFS на kartta.hel.fi; Earth Search
- Поиск ArcGIS и ArcGIS Hub; И. Бегтин, скрытые данные на сайтах ArcGIS, Telegram, 22.07.2024, поиск ArcGIS как каталог, 31.07.2024, и где искать геоданные, 07.06.2025 — оценки объёма
- И. Бегтин, геопортал Архангельской области после окончания лицензии ArcGIS, Telegram, 21.08.2024
- Росреестр, публичная кадастровая карта на платформе НСПД, CNews, 27.12.2024; И. Бегтин, портал пространственных данных Росреестра, Telegram, 07.01.2024
- Приказ Минэкономразвития России от 17.06.2021 № 349 о форматах ФГИС ТП (карты в GML и PDF); вопросы и ответы на сайте ФГИС ТП о выдаче векторной модели, снимок веб-архива от 26.05.2020; И. Бегтин, ФГИС ТП: скачать можно только растр, Telegram, 08.04.2024
- Постановление Правительства РФ от 04.03.2017 № 262 о предоставлении данных из фондов пространственных данных; И. Бегтин, когда открытые данные — не открытые, Telegram, 25.10.2021
- Доступность НСПД, ФГИС ТП, фонда пространственных данных Санкт-Петербурга и сайта Росреестра — проверка с немецкого IP-адреса, 12 октября 2026 года.

Научные репозитории данных

Научный репозиторий принимает данные исследования на хранение, закрепляет их версию, выдаёт постоянный идентификатор и даёт готовую ссылку для статьи. Задача одна — чтобы через годы по ссылке из статьи открылась та же версия тех же файлов.

От портала открытых данных его отличает не тема, а обязательства. Репозиторием каталог делают четыре вещи сразу:

- **приём на хранение** — автор передаёт файлы репозиторию, а не оставляет ссылку на свой сайт;
- **неизменная версия** — опубликованное не правят молча, исправление выходит новой версией;
- **постоянный идентификатор** — обычно DOI, который переживёт смену адреса и платформы;
- **описание для ссылки** — авторы, год, название, издатель и версия, из которых собирается строка цитирования.

Если чего-то нет, это файлохранилище или витрина, даже когда раздел сайта называется репозиторием.

От репозиториев ждут, что данные будут соответствовать принципам FAIR (2016 год): их можно найти, получить, совместить с другими и использовать повторно. FAIR не значит «открыто»: набор с закрытыми файлами тоже может быть FAIR, если его описание открыто, у него есть DOI и понятно, как запросить доступ.

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) научных репозиториев 8404 — 19% всех каталогов, второй по численности тип после геопорталов. Действуют 8285. Недействующих и выведенных из эксплуатации 119, около 1,4%: научные репозитории закрываются реже порталов открытых данных, у тех около 4%. Ограниченный доступ хотя бы к части данных отмечен у 527 репозиториев, 6%.

Международный реестр re3data 12 октября 2026 года описывал 3535 репозиториев. Разница с реестром каталогов не ошибка: re3data подробно описывает репозитории, которые прошли его редакцию, а реестр каталогов считает и университетские репозитории публикаций, где данных немного, и серверы данных вроде THREDDS и ERDDAP.

Зачем учёные публикуют данные

Учёный кладёт данные в репозиторий чаще всего потому, что этого требуют фонд или журнал.

- **Фонды.** В программе ЕС Horizon Europe план управления данными обязателен, а данные открывают по правилу «настолько открыто, насколько возможно, настолько закрыто, насколько необходимо». Национальные институты здоровья США (NIH) с 25 января 2023 года требуют план управления данными и обмена ими в каждой заявке, где появляются научные данные. Меморандум Управления научно-технической политики США 2022 года

обязал федеральные ведомства к концу 2025 года ввести политики, по которым публикации и данные федеральных исследований открываются без задержки. Большинство таких политик уже действует, но с февраля 2026 года администрация пересматривает сам меморандум.

- **Журналы.** Многие журналы требуют в статье заявление о доступности данных: где лежат данные и на каких условиях. PLOS ввела такое требование в 2014 году.
- **Признание.** На набор с DOI можно сослаться, и ссылки на него считаются. Статья о данных даёт за работу с данными научную публикацию, о ней — в разделе «Статьи о данных».

Заявление «данные по запросу у авторов» репозиторий не заменяет. В исследовании 2022 года авторам 1792 статей, которые обещали данные по запросу, написали с просьбой их прислать. 93% не ответили или отказали, данные прислали 123 автора — 6,8%.

Для каталога из этого следует, что DOI, долговременное хранение, лицензия и описание для ссылки здесь не дополнительные функции. Без них фонд не примет репозиторий в план управления данными. NIH, например, советует сначала искать предметный репозиторий своей дисциплины, а если его нет — универсальный или университетский. Список требований к репозиторию у NIH начинается с того же: постоянный идентификатор и страница набора, которая открывается, даже если сам набор изъят. Репозиторий сверяют по re3data и по сертификату CoreTrustSeal, о нём — в разделе «Стандарты».

Кто создаёт

Научные репозитории почти целиком ведут университеты и научные организации.

Владелец	Репозиториев	Доля
Университеты и научные организации	7 164	85%
Центральные органы власти	592	7%
Международные организации	217	3%
Общественные организации	137	2%
Компании	117	1%
Регионы	79	1%
Прочие: сообщества, местные и федеральные органы власти	98	1%

Центральные органы власти здесь — ведомственные научные службы: геологические, метеорологические, океанографические. Международные — консорциумы вроде GBIF и межправительственные организации.

По тому, кто может положить данные, репозитории делятся на пять видов.

Вид	Кто ведёт и для кого	Что даёт автору	Примеры
Институциональный	университет или институт для своих сотрудников	модерацию библиотеки, свой префикс DOI, часто — общий репозиторий со статьями и диссертациями	DaRUS (Штутгарт), архив СПбГУ
Предметный	научное сообщество для одной дисциплины	строгий приём по стандартам дисциплины, читателей из своей науки	PANGAEA, GBIF, ICPSR

Вид	Кто ведёт и для кого	Что даёт автору	Примеры
Универсальный	сервис для любой науки и любого автора	быструю публикацию с DOI без своей установки	Zenodo, Figshare, Harvard Dataverse, Dryad, Mendeley Data, OSF
Национальный или общий	государство или консорциум для всех организаций страны	готовый репозиторий без своего администратора	Fairdata (Финляндия), JAIRO Cloud (Япония), DABAR (Хорватия)
CRIS	университет как система учёта научной работы	набор данных как ещё один результат рядом со статьями и грантами	порталы на Pure, Research.fi

О национальных моделях — в разделе «Национальные модели», о CRIS — в разделе «CRIS — скрытые каталоги данных».

Число репозиториев — не мера научных данных

По числу научных репозиториев в реестре лидируют США (2018), Германия (764), Великобритания (503), Франция (289), Япония (275), Канада (245), Китай (244) и Италия (211). Ещё 681 репозиторий — международные. Как и с порталами открытых данных, счёт показывает не объём данных, а то, как устроена публикация в стране. В Японии 160 репозиториев из 275 — это витрины общего облачного сервиса JAIRO Cloud на программе WEKO3. В Хорватии 76 из 84 — репозитории организаций в национальной системе DABAR. В Германии 121 — серверы публикаций университетских библиотек на OPUS, в США 146 — университетские репозитории на облачном сервисе Digital Commons. В большинстве из них лежат статьи и диссертации, а данные — малая доля.

Что публикуют

Публикуют таблицы и файлы к статьям, наблюдения, код, описания коллекций. Хороший пакет набора — это не только файлы: в нём описание методики, словарь полей, код обработки и лицензия. Если репозиторий общий, рядом лежат и сами публикации. При сборе описаний статьи приходится отделять от данных, иначе поисковик данных покажет диссертации.

Как публикуют данные, сильно зависит от дисциплины. Науки о жизни, особенно биоинформатика, давно живут в своей связанной инфраструктуре: свои стандарты, идентификаторы и базы, где единица учёта — исследование или геном, а не набор. Науки о Земле — климат, метеорология, океан — тоже держатся особняком со своими форматами и серверами. Данные для машинного обучения в основном лежат на Kaggle и Hugging Face, о них — глава о ML-каталогах. Экономика, социальные и гуманитарные науки чаще кладут данные туда же, куда и статьи: в репозиторий университета или в общий репозиторий.

В реестре это видно по темам: после общей темы «наука и технологии» чаще всего встречаются окружающая среда (около 2,3 тысячи репозиториев), живые организмы (около 1,6 тысячи), здоровье, океаны, науки о Земле и климат. Естественные науки описаны и собраны лучше, чем социальные и гуманитарные: у них есть своё ПО и свои сети, о которых — раздел «Дисциплинарные инфраструктуры».

Стандарты

Описание. Основа — схема DataCite: последняя версия 4.7 вышла в марте 2026 года. Обязательных полей шесть: идентификатор, авторы, название, издатель, год публикации и тип ресурса. По протоколу OAI-PMH репозитории отдают ещё короткий Dublin Core. У дисциплин свои схемы: DDI для социальных обследований, Darwin Core и EML для биоразнообразия, ISO 19115 для наук о Земле, соглашения CF для файлов NetCDF. Платформы отстают от схемы: InvenioRDM по своей документации опирается на версию DataCite 4.3. Если вам нужны новые поля, например связь с проектом через RAiD, проверьте, какую версию схемы поддерживает платформа.

Идентификаторы. Набор получает DOI, автор — ORCID, организация — ROR. По этим ссылкам поисковики и агрегаторы связывают набор с людьми, организациями и статьями. В описании набора на Harvard Dataverse в разметке schema.org у автора стоит ссылка на ORCID, а у организации — на ROR. DOI, версии и строка цитирования здесь обязательны. Как они устроены, разобрано в главах об идентификаторах и о цитировании.

Передача описаний. Главный интерфейс научных репозиториев — OAI-PMH: в реестре он отмечен у 2068 репозиториев, около четверти. Примерно столько же отдают карту сайта, а выгрузку DCAT — около 200. Для Google Dataset Search на странице набора нужна разметка schema.org. Новее — FAIR Signposting: страница набора в заголовке ответа или в отдельном файле перечисляет ссылки на описание, файлы, лицензию и идентификаторы авторов. Так программа узнаёт главное о наборе, не разбирая HTML. Signposting есть в ядре Dataverse с версии 5.14, InvenioRDM с версии 12 и DSpace с версии 7.6. Подробнее о форматах выгрузки — в главе об экспорте метаданных.

Доверие. CoreTrustSeal — сертификат долговременного хранения по 16 требованиям: миссия, права, непрерывность доступа, хранение, безопасность. Сертифицированных репозиториев около 160, требования обновлены на 2026–2028 годы. В 2026 году совет CoreTrustSeal одобрил отдельную методику, как репозиторию готовиться к собственному закрытию. Сертификат и профиль в re3data — первое, что смотрят фонды и журналы, когда проверяют репозиторий.

Платформы

Самописных научных репозиториев около трети, и именно у них хуже всего с программным доступом. Среди готовых платформ много таких, что сделаны не для данных.

Платформа	В реестре	Что это	API отмечен
Своя разработка	2 836	—	5%
DSpace	829	свободная платформа библиотечных репозиториев: публикации и данные	99%
Pure	369	коммерческая CRIS компании Elsevier	100%
GBIF IPT	360	свободный инструмент публикации данных о биоразнообразии	91%
Figshare	288	облачный сервис: общий figshare.com и репозитории организаций	99%
Dataverse	257	свободный репозиторий данных	99%
THREDDS	249	сервер массивов NetCDF и HDF	98%

Платформа	В реестре	Что это	API отмечен
ERDDAP	175	сервер рядов и сеток наблюдений	99%
Digital Commons	174	облачный сервис Elsevier для репозиториев университетов, в основном в США	99%
WEKO3	161	программа японского сервиса JAIRO Cloud	100%
EPrints	129	свободная платформа репозиториев публикаций	99%
OPUS	124	свободный сервер публикаций немецких библиотек	98%

Реже встречаются InstDB (83, Китай), DABAR (76, Хорватия), InvenioRDM (74, включая Zenodo), DSpace-CRIS (44) и CKAN (39).

Из таблицы видно три вещи. Первая: платформы делятся на четыре семьи — репозитории публикаций (DSpace, EPrints, OPUS, Digital Commons, WEKO3), CRIS (Pure), репозитории данных (Dataverse, InvenioRDM, Figshare) и серверы данных (THREDDS, ERDDAP, IPT). Для наборов данных спроектирована только третья семья, поэтому в репозиториях на платформах первой семьи данных мало. Вторая: у готовых платформ API почти всегда есть, а у своих разработок он отмечен только у 5%. Довод против самописного репозитория здесь ещё сильнее, чем у порталов открытых данных. Третья: недействующих больше всего среди установок IPT — 12%, у остальных крупных платформ не больше 2%.

Что в ядре, а что нет

DOI есть в ядре Dataverse, InvenioRDM, Figshare и DSpace, готовая ссылка с выгрузкой в BibTeX и RIS — у первых трёх. В DSpace до версии 10.1 включительно есть только метатеги для Google Scholar, кнопку «Cite» переносят из DSpace-CRIS. Ставить ради науки портал открытых данных и прикручивать DOI расширением — обходной путь.

Возможность	Dataverse	InvenioRDM	DSpace	CKAN
DOI	ядро: DataCite, EZID, также Handle	ядро: DataCite	ядро, включается настройкой; по умолчанию Handle	расширение ckanext-doi, заявлены CKAN 2.9 и 2.10
Версии набора	ядро	ядро: свой DOI у каждой версии и общий на все	ядро	нет
Эмбарго и закрытые файлы	ядро, с запросом доступа	ядро, с запросом доступа	ядро: эмбарго до даты, запрос копии файла	только закрытый набор целиком
OAI-PMH	ядро	ядро	ядро	нет в ядре
Разметка schema.org	ядро	ядро	в ядре не найдена	расширение ckanext-dcat
DCAT	нет	ядро (DCAT-AP)	нет	расширение ckanext-dcat
FAIR Signposting	ядро с 5.14	ядро с v12	ядро с 7.6	нет
Строка цитирования, BibTeX, RIS	ядро	ядро (стили CSL, BibTeX, CSL-JSON; RIS нет)	нет: только метатеги для Google Scholar	расширение ckanext-doi (только при DOI) или ckanext-citeproc
Предметные поля описания	ядро: готовые блоки и свои	настройка ядра	настройка ядра	расширение ckanext-scheming

Расширения у научных платформ устроены иначе, чем у CKAN: основное в ядре, а снаружи — просмотрщики файлов и дополнительные форматы выгрузки. У Dataverse это внешние инструменты и внешние экспортёры, их ведёт сообщество GDCC, например экспортёры RO-Crate и DCAT 3; экспортёр Croissant с версии 6.10 перенесён в ядро. Сама документация Dataverse называет механизм экспортёров экспериментальным и предупреждает, что при обновлении платформы он может потребовать доработки. У DSpace вместо плагинов — модели сущностей и настройки: в DSpace 10 (2026 год) появился отдельный тип записи «набор данных» со своей страницей и подробным списком файлов. CKAN для науки остаётся обходным путём: расширение для DOI текущую линейку ядра документацией не подтверждает, подробности в главе про CKAN.

Версии и сопровождение. На 12 октября 2026 года последние выпуски: Dataverse 6.12.1 (8 октября), InvenioRDM v14 (7 августа; v13 поддерживается до 7 февраля 2027 года, v15 в бете), DSpace 10.1 (14 сентября; в тот же день вышли исправления 9.4 и 8.5). Крупные установки отстают на одну-две версии: Harvard Dataverse работает на 6.10.1, DSpace@MIT — на 8.2. Это нормально, если обновления запланированы. Сколько стоит сопровождение, разобрано в главе о развёртывании, карточки платформ — в части «Платформы и расширения».

Сколько можно положить в универсальный репозиторий

Универсальные репозитории бесплатны до предела, а большие данные стоят денег.

Сервис	Бесплатно	Сверх лимита
Zenodo	до 50 ГБ и 100 файлов на запись, число записей не ограничено	один раз до 200 ГБ на запись по запросу. Дробить один набор на записи ради обхода лимита правила запрещают
Harvard Dataverse	файлы до 2,5 ГБ через браузер, до 1 ТБ на исследователя	крупнее — по договорённости, прежде всего для исследователей Гарварда
Figshare	20 ГБ личного места, файл до 20 ГБ	Figshare+: разовая плата 1225 долларов за каждые 250 ГБ, файлы до 5 ТБ
Dryad	авторам из организаций-партнёров — до 10 ГБ за счёт партнёра	остальным 150 долларов за набор до 5 ГБ и дальше по шкале, до 12 162 долларов за 2 ТБ

Лимиты и цены — на 12 октября 2026 года, у Dryad тарифы действуют с 6 мая 2025 года. Если данных терабайты, универсальный репозиторий не подходит: нужен предметный архив дисциплины или своё хранилище с описанием в каталоге. Варианты — в главах о хранилище данных и об облачных хранилищах и хабах.

Дисциплинарные инфраструктуры

У многих наук есть свои каталоги на своём ПО. Это полноценные каталоги: с описаниями, лицензиями и программным доступом. Но за пределами дисциплины о них мало кто знает, и общие поисковики данных их почти не собирают.

- **ERDDAP** — сервер данных Национального управления океанических и атмосферных исследований США (NOAA) для наблюдений за океаном, погодой и климатом. Ряды и сетки он отдаёт через API в CSV, JSON, NetCDF и других форматах. По сайту проекта, ERDDAP поставили около 100 организаций как минимум в 17 странах, в реестре около 175 серверов.

- **THREDDS Data Server** от Unidata и серверы **OPeNDAP Hyrax** раздают большие массивы в NetCDF и HDF. В реестре около 250 серверов THREDDS и около 100 серверов OPeNDAP. Описания у них скудные: снаружи такой сервер больше похож на научное файлохранилище, чем на каталог.
- **GBIF IPT** — инструмент публикации данных о биоразнообразии. Его ставят университеты и музеи, а глобальная сеть GBIF собирает данные со всех установок. В октябре 2026 года в GBIF зарегистрировано 392 установки IPT, в самой сети около 125 тысяч наборов и почти 4 млрд записей о встречах видов. Каждая установка отдаёт описания в DCAT. В России так работает, например, IPT Зоологического института РАН (ipt.zin.ru).
- **WIS 2.0** — система Всемирной метеорологической организации для обмена данными метеостанций. Открытый пакет wis2box отдаёт данные по стандарту OGC API. В реестре около 95 установок, среди них российская (wis2box.mecom.ru) и казахстанская. Предшественник — OpenWIS, сделанный на основе GeoNetwork. Его код не развивается с 2022 года, в реестре осталось 11 таких порталов.

Вывод для владельца данных такой дисциплины: ставьте её стандартный инструмент, а не общий портал. Так данные попадут в сеть, где их ищут коллеги. Чтобы их нашли и за её пределами, нужен ещё экспорт описаний в DCAT или schema.org, об этом глава о находимости.

Национальные модели

Страны по-разному собирают научные данные в одну систему.

- **Финляндия.** Министерство образования и культуры предлагает набор сервисов Fairdata, которые делает центр CSC: хранилище IDA, поиск наборов Etsin, редактор описаний Qvain, интерфейс метаданных Metax, долговременное хранение и платформу публикации AVAA. Учёный проходит путь от хранения до публикации в одной системе.
- **Япония.** С 2012 года Национальный институт информатики (NII) даёт университетам облачный репозиторий JAIRO Cloud на своей программе WEKO3. С 2016 года сервис ведут вместе с консорциумом открытых репозиториев JPCOAR, с 2017 года он платный, 100 ГБ входят бесплатно. На конец 2025 финансового года им пользовались 822 организации. Менять программу под себя организация не может, зато обновления, серверы и выгрузку описаний берёт на себя NII. Модель выросла из нехватки людей: по данным NII, почти у 90% членов консорциума репозиторием занимается меньше одного сотрудника на полную ставку. В 2026 году в сервис добавляют запрос доступа к закрытым файлам.
- **Хорватия.** Национальная система DABAR с 2015 года, её ведёт вычислительный центр SRCE Загребского университета. В ней 185 репозиториев научных организаций и вузов. В ноябре 2025 года систему перевели на новую архитектуру на основе Fedora Commons и связали с национальной CRIS CroRIS. Открытой при этом, по исследованию 2025 года, была лишь чуть больше половины объектов.
- **Германия.** Единой платформы нет, зато есть набор открытых продуктов. Dataverse — DaRUS в Штутгарте, heiDATA в Гейдельберге, в реестре 25 немецких установок. Invenio — архив RODARE. DSpace — OpARA в Дрездене. Местный MyCoRe — в реестре 22 немецких

портала, например opendata.uni-kiel.de. Рядом большие предметные архивы вроде PANGAEA и агентство da|ra, которое выдаёт DOI социальным и экономическим данным и ведёт по ним поиск.

- **Китай.** Science Data Bank (ScienceDB, scidb.cn) — общий репозиторий, который ведёт Компьютерный сетевой информационный центр Академии наук Китая. Институты публикуют данные на типовом ПО InstDB из пакета FairStack, в реестре 82 китайских установки. InstDB поддерживает три режима доступа: открытый, после авторизации и по запросу. Поверх работает поисковик FinData (findata.cn). Снаружи эта система почти не видна: в декабре 2025 года из 72 известных установок InstDB с зарубежных адресов открывались от 6 до 12.
- **США.** Данные собирают ведомственные агрегаторы. Каталог данных Геологической службы (USGS) объединяет несколько репозиториев, главный из них ScienceBase. По наблюдению 2022 года, DOI, привязка к месту и период наблюдений там обязательны для каждого набора. Министерство энергетики ведёт OSTI.GOV и DOE Data Explorer, а Science.gov ищет по научным результатам многих федеральных ведомств, включая данные.
- **Франция.** Национальный репозиторий Recherche Data Gouv работает на Dataverse.
- **Россия.** Национальной инфраструктуры научных данных нет. По подсчёту июля 2025 года, наборов с DOI DataCite и российской аффилиацией было 6806, большая часть — от Российской академии наук, и почти все загружены на зарубежные платформы вроде Zenodo, Dryad и Mendeley Data. В реестре 75 российских научных репозиториев, больше половины — самописные. Где наука встроена в мировые сети, данные публикуют на их инструментах, как в узлах GBIF IPT.

Японская и хорватская модели — то же, что типовой портал у открытых данных: одна сборка для сотен организаций, у которых нет своих людей. Работает это, когда общая система сама выдаёт DOI и отдаёт описания наверх, а организации остаётся только наполнять её. Слабое место то же, что у типового портала: в такие системы кладут прежде всего статьи и диссертации, а данные остаются малой долей.

CRIS — скрытые каталоги данных

CRIS (current research information system) — система учёта научной работы организации: публикации, проекты, гранты, сотрудники, оборудование. Многие университеты публикуют данные не в отдельном репозитории, а в CRIS, как ещё один результат исследования. Самая распространённая система — Elsevier Pure, в реестре около 370 порталов на ней. Рядом DSpace-CRIS (около 45) и Converis. В Pure набор данных — отдельный тип записи. По умолчанию он выключен, его включает администратор, а DOI можно выпускать прямо из системы через DataCite. Национальный пример — финский Research.fi: публикации, проекты, финансирование, исследователи и наборы данных страны в одном поиске.

Скрытыми эти каталоги делает то, что портал выглядит как сайт университета о науке, а не как каталог данных, и наборы теряются среди статей и профилей. Коммерческие платформы к тому же закрываются от программ: 12 октября 2026 года четыре университетских портала

на Pure на автоматический запрос ответили отказом или проверкой на робота. Чтобы данные из CRIS находили, у набора должен быть DOI DataCite: тогда его описание живёт и вне сайта университета.

Почему их трудно собирать

Научных данных много, но собрать их в один поиск труднее, чем порталы открытых данных или статистику. Причины технические и коммерческие.

1. **Протокол OAI-PMH не про файлы.** Это главный интерфейс научных репозиториев. Обязательный формат в нём — простой Dublin Core, где нет места для списка файлов: обычно в записи есть только ссылка на карточку. Отбор по разделам (sets) задаёт сам репозиторий, и попросить у него «только данные» протокол не позволяет.
2. **Данных мало среди статей.** В DSpace фильтр по типу записи описан в настройках, но в поиск по умолчанию не включён. Чтобы найти наборы, сборщик скачивает описания всех записей и отбирает данные сам. В архиве открытого доступа СПбГУ в декабре 2024 года из 47 619 записей наборами данных были помечены 17. В октябре 2026 года архив работает на DSpace 10.1, и фильтра по типу записи среди фильтров поиска по-прежнему нет. По оценке того же года, в репозиториях на DSpace наборы данных — меньше 5% записей.
3. **Коммерческие платформы.** Figshare и порталы на Pure составляли около трети научных каталогов в реестре Dateno в 2024 году. Автоматический сбор они ограничивают.
4. **Файлы лежат в другом месте.** Всё чаще учёный кладёт файлы в Zenodo или Mendeley Data, а в репозитории университета остаётся карточка со ссылкой на чужую карточку.
5. **Агрегаторы считают данными всё подряд.** В DataCite в октябре 2026 года больше 76 млн записей с типом «набор данных», среди них нарезки больших баз и объекты, далёкие от таблиц. Подробнее — в главе «Мир каталогов в цифрах».

Что из этого следует для владельца репозитория: включить фильтр по типу записи или отдельную коллекцию для данных; если платформа умеет, отдавать по OAI-PMH не только Dublin Core, но и схему DataCite; хранить файлы у себя или давать прямые ссылки на них; включить FAIR Signposting; завести профиль в re3data.

Статьи о данных

Статья о данных (data paper) — рецензируемая статья, которая описывает набор: как его собрали, что в нём есть, как им пользоваться. Гипотез и выводов в ней нет. Сами данные лежат в репозитории, статья ссылается на них и даёт авторам научное признание за работу с данными. Такие статьи печатают Scientific Data, Earth System Science Data, Biodiversity Data Journal, Data in Brief и другие журналы. GBIF ведёт их список, а IPT умеет сделать из описания набора черновик статьи. Пока их немного: в DataCite в октябре 2026 года около 18,5 тысячи записей с типом «статья о данных».

Для каталога вывод простой: если к набору есть статья, свяжите их, указав DOI статьи в описании набора, и наоборот. Для поисковика данных вопрос открыт. Статья о данных — не сами данные, но иногда единственное место, где о наборе вообще сказано.

Примеры

Универсальные

- Zenodo — общий репозиторий при ЦЕРН на InvenioRDM. Принимает данные, код и публикации любой дисциплины, DOI у каждой версии.
- Harvard Dataverse — установка Dataverse при Гарвардском университете, бесплатно открыта исследователям всего мира.
- Figshare — облачный сервис: общий бесплатный репозиторий и платные репозитории университетов.
- Dryad — некоммерческий репозиторий с проверкой каждого набора куратором; платят организации-партнёры или авторы.

Предметные

- PANGAEA — науки о Земле и среде.
- GBIF — глобальная сеть данных о биоразнообразии, которая собирает наборы с сотен установок IPT.
- ICPSR — архив данных социальных наук при Мичиганском университете.

Национальные и общие

- Etsin — поиск научных наборов Финляндии в системе Fairdata.
- JAIRO Cloud — общий облачный репозиторий японских университетов.
- DABAR — национальная система репозиториев Хорватии.
- Science Data Bank — общий репозиторий Академии наук Китая.
- Recherche Data Gouv — национальный репозиторий Франции на Dataverse.

Институциональные

- CaltechDATA — репозиторий Калифорнийского технологического института на InvenioRDM, около 77 тысяч записей в октябре 2026 года.
- DaRUS — репозиторий данных Штутгартского университета на Dataverse.

Где ломается

Репозиторий публикаций под видом репозитория данных. Каталог называют репозиторием данных, а данных в нём почти нет: в архиве СПбГУ 17 наборов на 47 тысяч записей. Сборщик не может попросить «только данные» и забирает тысячи статей. Почему так выходит, разобрано в разделе «Почему их трудно собирать».

DOI есть, файла нет. DOI ведёт на карточку, а файл удалён, лежит на чужом сайте или выдаётся «по запросу». Требование NIH здесь точное: страница набора должна открываться, даже если набор изъят, и объяснять, что с ним стало. Пустая страница с ошибкой ломает ссылку из статьи.

Правят опубликованное. Файл заменили без новой версии, и тот, кто сослался на набор, теперь ссылается на другие данные. Научные платформы этому мешают: Zenodo разрешает менять файлы опубликованной записи только 30 дней, дальше — новая версия или

обоснованный запрос в поддержку. На портале открытых данных файл обычно меняют молча, и для науки это поломка.

Репозиторий закрылся. По исследованию 2023 года, из 3069 репозиториев в re3data закрылись 191, или 6,2%. Средний возраст на момент закрытия — 12 лет. Только 44% закрытых передали данные в другой репозиторий, и почти в половине случаев данные, возможно, потеряны. Отсюда требования к плану закрытия у CoreTrustSeal и выводы в главе об архивации.

Робот не проходит. 12 октября 2026 года страница набора на Harvard Dataverse на автоматический запрос ответила не самой страницей, а проверкой на робота от облачного фильтра AWS, хотя API и OAI-PMH отвечали нормально. Порталы на Pure в тот же день отвечали отказом. Для поисковика или сборщика, который читает страницу набора, разметку schema.org и заголовок Signposting, такой набор не существует. Защита от ботов нужна, но машинные входы — страница набора, API, OAI-PMH — должны оставаться открытыми.

Свой репозиторий без API. У самописных научных репозиториев программный интерфейс отмечен только у 5%. Такой репозиторий не соберут ни научные агрегаторы, ни поисковики данных.

Ставка на одного человека. Репозиторий живёт, пока его ведёт конкретный сотрудник или аспирант. У установок IPT доля недействующих — 12%, в несколько раз выше, чем у других крупных платформ. Общие сервисы вроде JAIRO Cloud и DABAR появились именно как ответ на это.

Как проверить научный репозиторий

Пять вопросов, на которые можно ответить за полчаса.

1. **Есть ли у набора DOI и куда он ведёт?** Запрос к DataCite покажет адрес страницы, тип записи и версию.
2. **Отдаёт ли репозиторий описания сборщикам?** Протокол OAI-PMH должен отвечать, а среди форматов, кроме простого Dublin Core, должна быть схема DataCite (`oai_datacite`).
3. **Можно ли отобрать только данные?** Посмотрите разделы OAI-PMH и фильтры поиска: есть ли отдельная коллекция или тип записи «набор данных».
4. **Видят ли набор программы?** Проверьте разметку schema.org и заголовок `Link` со ссылкой на `linkset` — это FAIR Signposting. Если вместо страницы приходит проверка на робота, поисковики набор не прочитают.
5. **Числится ли репозиторий в re3data и есть ли у него сертификат?** Это то, что проверит фонд.

```
# DataCite: куда ведёт DOI, тип записи и версия
curl -s "https://api.datacite.org/dois/10.7910/DVN/TJCLKP"
# OAI-PMH: жив ли интерфейс, какие схемы он отдаёт и на какие разделы делится
curl -s "https://dataverse.harvard.edu/oai?verb=Identify"
curl -s "https://dataverse.harvard.edu/oai?verb=ListMetadataFormats"
curl -s "https://dataverse.harvard.edu/oai?verb=ListSets"
# Dataverse: версия платформы и описание набора в schema.org
curl -s "https://dataverse.harvard.edu/api/info/version"
```

```
curl -s "https://dataverse.harvard.edu/api/datasets/export?
exporter=schema.org&persistentId=doi:10.7910/DVN/TJCLKP"
# InvenioRDM: число записей и заголовок FAIR Signposting у страницы записи
curl -s "https://data.caltech.edu/api/records?size=1"
curl -sI "https://data.caltech.edu/records/0jxb3-ny611" | grep -i '^link'
# DSpace 7 и новее: версия платформы
curl -s "https://dspace.mit.edu/server/api"
# re3data: есть ли репозиторий в реестре
curl -s "https://www.re3data.org/api/beta/repositories?query=PANGAEA"
```

Все команды проверены 12 октября 2026 года. Harvard Dataverse отдаёт по OAI-PMH, кроме Dublin Core, схемы DataCite и DDI. Страница записи CaltechDATA возвращает заголовок со ссылкой на `linkset`. Подробный список проверок — в чек-листах, команды для идентификаторов — в главе об идентификаторах.

Когда свой репозиторий не нужен

Свой репозиторий — это постоянная работа: обновления, модерация, договор с DataCite, ответы авторам. Небольшой группе часто достаточно Zenodo или предметного репозитория своей дисциплины, и фонды это принимают. Университету без людей в библиотеке выгоднее подключиться к общему сервису страны или консорциума, как японские и хорватские вузы. Свой Dataverse или InvenioRDM оправдан, когда библиотека модерирует десятки групп, нужен свой префикс DOI и есть кому сопровождать систему. Разбор такого решения — в примере университета. Требования для научной организации — в главе о сценариях, для консорциума — в соседнем разделе.

Чем отличается от соседей

Сосед	Главное отличие
Портал открытых данных	важнее регулярное обновление, чем неизменная версия с DOI; публикует орган власти, а не автор исследования
Каталог микроданных	данные о людях из обследований, доступ по заявке и описание в DDI. Архивы социальных наук вроде ICPSR стоят на стыке двух типов
ML-каталог	набор нужен, чтобы загрузить его в обучающий код, а не чтобы сослаться в статье; рядом карточка набора и разбиение на части
Геопортал	главный продукт — слой и сервис, а не зафиксированный файл. Серверы THREDDS и ERDDAP стоят на стыке
Репозиторий публикаций	единица — статья или диссертация, данные — малая доля записей
CRIS	учёт всей научной работы организации, набор данных — лишь один из типов записей

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: владельцы, страны, платформы с долей API и недействующих, главные платформы по странам, интерфейсы, темы и доступ научных репозиториев. Запросы — в приложении.
- re3data и его API — число репозиториев, 12 октября 2026 года

- M. D. Wilkinson и др., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 2016; принципы FAIR на GO FAIR
- Политики фондов и журналов: открытая наука в Horizon Europe; политика NIH об управлении данными и обмене ими, действует с 25 января 2023 года, включая требования к репозиториям; выбор репозитория по NIH; политики федеральных ведомств США по меморандуму OSTP 2022 года, SPARC, обновлено 4 мая 2026 года; о пересмотре меморандума, STM, 19 февраля 2026 года; требования PLOS к доступности данных
- M. Gabelica, R. Bojčić, L. Puljak, Many researchers were not compliant with their published data sharing statement, Journal of Clinical Epidemiology, 2022
- Схема метаданных DataCite, версия 4.7 от 3 марта 2026 года; описание метаданных InvenioRDM — поддерживаемая версия схемы
- FAIR Signposting и кто его поддерживает; Signposting в Dataverse, InvenioRDM и DSpace
- CoreTrustSeal — требования 2026–2028 годов, число сертифицированных репозиториев и методика на случай закрытия, 12 октября 2026 года
- Платформы: форматы выгрузки Dataverse и внешние экспортёры; каталог экспортёров GDCC; форматы выгрузки InvenioRDM в исходном коде; политика поддержки InvenioRDM; DOI в DSpace; тип записи «набор данных» в DSpace 10; возможности Dataverse; выпуски Dataverse и DSpace. Версии Harvard Dataverse, DSpace@MIT и ответы OAI-PMH, Signposting и проверки на робота — проверка 12 октября 2026 года.
- Лимиты и цены: Zenodo и правка файлов после публикации; Harvard Dataverse; Figshare и Figshare+; Dryad — проверено 12 октября 2026 года
- Спецификация OAI-PMH 2.0 — обязательный формат Dublin Core и разделы; настройки поиска DSpace по умолчанию — фильтр по типу записи описан, но не включён; тип записи «набор данных» в Pure — проверено 12 октября 2026 года
- ERDDAP — кто использует; THREDDS Data Server; OPeNDAP; GBIF IPT и API реестра установок GBIF; wis2box; OpenWIS — проверено 12 октября 2026 года
- Национальные модели: Fairdata, Research.fi; JAIRO Cloud у JPCOAR — история, условия и 822 организации на конец 2025 финансового года; JAIRO Cloud у NII; доклад NII для COAR, 2026, — нагрузка на сотрудников; запрос доступа в JAIRO Cloud, 23 марта 2026 года; DABAR и о новой версии DABAR, 15 мая 2026 года; National Repository Infrastructure and Open Access Challenges: The Croatian Perspective, LIBER Quarterly, 2025; DaRUS, heiDATA, RODARE, OpARA, MyCoRe, da|ra, ScienceDB, FinData, каталог данных USGS, DOE Data Explorer, Science.gov, Recherche Data Gouv — проверено 12 октября 2026 года
- D. Strecker, H. Pampel, R. Schabinger, N. L. Weisweiler, Disappearing repositories: Taking an infrastructure perspective on the long-term availability of research data, Quantitative Science Studies, 2023
- Статьи о данных в GBIF; API DataCite — число записей по типам, 12 октября 2026 года
- И. Бегтин, Telegram: Германия, 04.04.2023; Fairdata, 30.07.2023; статьи о данных, 07.04.2024; InstDB в Китае, 07.04.2024; данные зависят от дисциплины, 12.09.2024; почему научные данные трудно индексировать, 16.11.2024; ERDDAP, 09.12.2024; OpenWIS, 23.12.2024; DSpace СПбГУ, 25.12.2024; IPT в России, 26.12.2024; научные каталоги Великобритании, 16.06.2025; DataCite и наборы по странам, 20.07.2025; доступность InstDB из-за рубежа, 25.12.2025; каталог данных USGS, 17.02.2022

Каталоги показателей

Каталог показателей отдаёт статистические ряды и таблицы с разрезами: время, территория, пол, отрасль. Человек не скачивает «файл за 2019 год», а выбирает разрез. Программа должна уметь то же самое запросом.

Так публикуют официальную статистику, макроэкономику, данные центральных банков, медицинскую статистику и показатели целей устойчивого развития. По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) таких каталогов около 3,4 тысячи, 8% всех каталогов. Действуют почти все: недействующих меньше 1%.

Главный вывод главы: каталог показателей полезен тогда, когда данные в нём первичны, а сайт с графиками — только одна из витрин. Вся база выгружается целиком, у каждого ряда есть постоянный адрес, справочники общие для всех таблиц. Конструктор таблиц в браузере без API и выгрузки оставляет данные на сайте, даже если сайт выглядит современно.

Как устроен показатель

В каталоге показателей другая единица, чем в портале открытых данных. Там главное — файл. Здесь — число, у которого есть координаты.

- **Показатель** — что измеряют: численность населения, курс валюты, уровень безработицы.
- **Разрез** (в стандартах — измерение) — по какому признаку делят: год, регион, пол, возраст, отрасль.
- **Справочник** — список допустимых значений разреза с кодами: коды регионов, валют, видов деятельности.
- **Значение** (наблюдение) — одно число на пересечении разрезов: население Франции на 1 января такого-то года.
- **Временной ряд** — значения одного показателя при одних и тех же разрезах, расположенные по времени.
- **Атрибуты** — пометки к значению: единица измерения, «предварительные данные», «разрыв ряда из-за смены методики».
- **Методика** (справочные метаданные, паспорт показателя) — как показатель считают, откуда берут исходные данные, когда методика менялась.

Все сочетания разрезов вместе образуют таблицу, которую в стандартах называют кубом. Каталог показателей — это поиск по кубам, способ вырезать из них нужную часть и выдать её человеку или программе.

Как это выглядит вживую, видно на курсе доллара к евро в портале данных Европейского центрального банка. У ряда есть ключ `EXR.D.USD.EUR.SP00.A`, и каждая его часть — значение одного разреза:

Часть ключа	Разрез	Значение
`EXR`	набор	курсы валют
`D`	частота	ежедневно

Часть ключа	Разрез	Значение
`USD`	валюта	доллар США
`EUR`	к какой валюте	евро
`SP00`	вид курса	справочный курс ЕЦБ
`A`	вариант ряда	стандартное значение

Тот же ключ работает и как адрес. Запрос к API ЕЦБ возвращает 12 октября 2026 года последнее значение (1,1206 за 9 октября) вместе с атрибутами: статус наблюдения, число знаков после запятой, единица, полное название ряда. Заменив `USD` на `JPY`, получаешь курс иены — без поиска нового файла. Так устроены все хорошие каталоги показателей: адрес ряда собирается из кодов справочников.

Что получает свою страницу

Свою страницу или адрес в каталоге показателей получают три уровня:

1. **Тема или база** — «Население», «Цены», «Платёжный баланс».
2. **Таблица** (в SDMX — поток данных, dataflow) — куб с фиксированным набором разрезов. Именно её обычно называют набором данных.
3. **Ряд** — одна линия на графике, на которую можно сослаться.

Поэтому число «наборов» в каталогах показателей нельзя сравнивать с порталами открытых данных. 12 октября 2026 года у Статистики Швеции было 2814 таблиц, у Статистики Норвегии — 3753, у Статистики Канады через API доступно 8270 таблиц, а у ЕЦБ всего 108 наборов — но внутри них миллионы рядов. Одна таблица заменяет сотни файлов «по годам и регионам» на обычном портале.

Кто создаёт

Почти половину каталогов показателей ведут центральные органы власти: статистические службы, центральные банки, министерства.

Владелец	Каталогов	Доля
Центральные органы власти	1 556	46%
Регионы	662	19%
Международные организации	367	11%
Местные органы власти	273	8%
Общественные организации	250	7%
Компании	140	4%
Университеты и научные организации	134	4%

Картина обратна порталам открытых данных, где почти половина у муниципалитетов. Статистику считают централизованно, и у города своих рядов немного.

- **Национальные статистические службы и центральные банки** — основа типа. Рядом министерства со своей отраслевой статистикой, прежде всего здравоохранения.

- **Международные организации** — Евростат, ОЭСР, Всемирный банк, МВФ, Банк международных расчётов, агентства ООН. Они не только публикуют свои базы, но и собирают ряды стран, поэтому им нужен обмен по стандарту.
- **Регионы** — статистические ведомства земель Германии, автономных сообществ Испании, штатов. Около 43% региональных витрин в реестре — собственные разработки.
- **Местные органы власти** — в основном нидерландские и фламандские муниципалитеты: 131 из 273 местных витрин работают на одной платформе Swing.
- **Общественные и научные проекты** собирают чужие ряды в один поиск. Так устроены Our World in Data и DBnomics французского исследовательского центра CEPREMAP: в октябре 2026 года DBnomics собирал ряды из 94 источников, от статистических служб до центральных банков.

По странам первые места — США (333), Нидерланды (186), Германия (152), Франция (103) и Испания (82); ещё 373 каталога охватывают весь мир. Как и у порталов открытых данных, число говорит об устройстве, а не о качестве статистики. В Нидерландах 167 каталогов — витрины Swing городов и провинций. В США 69 — сайты показателей здоровья округов и больничных сетей на одном облачном сервисе. Во Франции 38 — геостатистические обсерватории на платформе Géoclip.

Что публикуют

Макроэкономику, демографию, цены, рынок труда, социальную статистику, финансы, внешнюю торговлю. Отдельные пласты — медицинская статистика министерств здравоохранения, показатели целей устойчивого развития и местные показатели: население кварталов, занятость, жильё.

Статистику публикуют на одной из трёх ступеней. На первой это документы: сборники и бюллетени в PDF, из которых ряд за десять лет собирают руками. На второй — таблицы: Excel на сайте или конструктор, который строит таблицу по запросу и даёт её скачать. На третьей данные первичны, а сайт с графиками — одна из витрин: вся база выгружается целиком, у каждого показателя есть адрес в API, справочники общие. Каталог показателей в смысле этой главы начинается со второй ступени, а пользы от него много только на третьей.

Чего в каталоге показателей обычно нет: записей об отдельных людях и предприятиях (это каталог микроданных), исходных анкет, геометрии территорий. Границы регионов для карты приходится брать на геопортале, и коды в справочнике показателей должны совпадать с кодами слоя — иначе ряд не ложится на карту.

Стандарты

SDMX — главный стандарт обмена статистикой. Его ведут семь международных организаций: Банк международных расчётов, ЕЦБ, Евростат, МВФ, ОЭСР, ООН и Всемирный банк. SDMX — не один формат, а три части:

- **модель** — как описать куб: какие у таблицы разрезы, какие справочники к ним привязаны, какие атрибуты допустимы. Эти описания называются структурами;
- **форматы** — SDMX-ML (XML), SDMX-JSON и SDMX-CSV;

- **API** — правила адресов, по которым программа получает данные и структуры: список таблиц, справочники, срез ряда.

Действующие сервисы чаще всего работают по версии 2.1: на ней API ЕЦБ, Евростата и МВФ. Версия 3.0 вышла в 2021 году, 3.1 — в мае 2025 года; 3.1 позволяет добавлять к таблице разрезы, не ломая уже собранные данные. Для общих тем есть глобальные структуры: национальные счета, платёжный баланс, цели устойчивого развития. Если служба передаёт данные по такой структуре, международная организация забирает их без ручной перекодировки. В реестре адреса SDMX отмечены у 177 каталогов показателей — около 5%.

JSON-stat — лёгкий формат куба в JSON: значения, разрезы и подписи в одном ответе. Версия 2.0 — формат по умолчанию у API Евростата и у нового API PxWeb. Программисту проще SDMX, но обмена справочниками и структурами между организациями в нём нет.

PC-Axis (PX) — файловый формат таблиц со встроенным описанием, родной для PxWeb. Его же отдают многие порталы открытых данных: в 2023 году испанский datos.gob.es содержал больше 24 тысяч наборов в PX, ирландский data.gov.ie — больше 7 тысяч.

Методика. Описание показателя — как считают, единица измерения, разрывы ряда — хранят отдельно от чисел, но рядом с ними. Без него ряд нельзя сравнивать между годами. В Европейской статистической системе методику описывают по единой структуре SIMS; МВФ требует описаний к данным, которые страны публикуют по его стандартам распространения.

Стандарты МВФ. Страны, которые подписались под стандартами распространения данных МВФ (e-GDDS, SDDS, SDDS Plus), публикуют национальную страницу сводных данных — типовую страницу с главными макроэкономическими рядами, графиком выпусков и часто с файлами SDMX. В реестре 118 таких страниц. Это не каталог в полном смысле, а обязательный минимум, по которому видно, что и когда страна публикует.

DCAT для статистики. Чтобы таблицы статистической службы появились на портале открытых данных, их описания нужно отдать в DCAT. Для этого есть профиль StatDCAT-AP: он добавляет к DCAT-AP разрезы, атрибуты и ссылку на структуру SDMX. Действующая версия — 1.0.1 (2019). Версия 3.0.0 приведена к DCAT-AP 3.0.1 и с 10 июня 2026 года остаётся рабочим черновиком. На деле выгрузку DCAT в реестре удалось найти меньше чем у десятка каталогов показателей. Поэтому статистика плохо видна на порталах открытых данных и в поисковиках данных, если служба не выкладывает её туда отдельно.

Связанные данные. У W3C есть словарь RDF Data Cube (2014) для статистики в виде связанных данных. Встречается редко, в основном в исследовательских проектах.

Платформы

Самописных витрин 44% (1505) — больше, чем у любого крупного типа. Программный доступ есть только у 37% каталогов показателей, хуже, чем у порталов открытых данных (69%), геопорталов (70%) и научных репозиториев (63%). У самописных витрин API отмечен лишь у 8%.

Платформа	Каталогов в реестре	Что это	Программный доступ
Своя разработка	1 505	—	у 8%

Платформа	Каталогов в реестре	Что это	Программный доступ
Swing	214	коммерческая платформа показателей ABF Research для муниципалитетов и провинций Нидерландов и Фландрии	отмечен у 91%
WordPress, Drupal	134 и 131	сайты со страницами показателей	API самой CMS — страницы и записи, а не ряды
Knoema	125	облачный сервис; в основном африканские витрины Open Data for Africa	сервис свёрнут, см. «Где ломается»
Национальная страница сводных данных МВФ	118	типовая страница по стандартам МВФ	чаще всего файлы SDMX
PxWeb	110	свободная платформа Статистики Швеции	свой API, JSON-stat, CSV, PX
DHIS2	74	свободная система медицинской статистики Университета Осло	Web API в ядре
OpenSDG	72	свободный статический сайт целей устойчивого развития	выгрузка всех показателей файлами
Conduent HCI	68	облачный сервис показателей здоровья для округов и больниц США	нет
Géoclip	53	коммерческая платформа геостатистических обсерваторий, в основном Франция	нет
.Stat Suite	40	свободная платформа SDMX сообщества SIS-CC	SDMX API в ядре
SDMX-RI	36	свободные компоненты SDMX Евростата	SDMX API
Power BI, R Shiny, Tableau	42, 39 и 29	дашборды	почти нет: у 0–5%

Видно, что тип делится на три мира. Первый — специализированное ПО статистики, где API и выгрузка есть из коробки: PxWeb, .Stat Suite, SDMX-RI, DHIS2, OpenSDG. Второй — облачные сервисы для местных и отраслевых витрин: Swing, Géoclip, HCI, Knoema; удобно, но данные живут у поставщика. Третий — сайты на CMS, дашборды и самописные порталы, где показатель есть на экране, но не у программы.

Тиражируемое ПО статистики

Программ, которые ставят у себя многие статистические службы, мало.

- **PxWeb** разработала Статистика Швеции, на нём работают службы скандинавских стран и десятки других, от Эстонии до Филиппин. Последний выпуск старой линии — 2025.v1 (февраль 2025 года). Следующее поколение — два отдельных приложения. API PxWebApi 2 Швеция сделала вместе со Статистикой Норвегии и запустила в октябре 2025 года (выпуск 2.7.0 — сентябрь 2026 года). Новый интерфейс PxWeb 2 берёт данные из этого API (выпуск 2.12.0 — 7 октября 2026 года). Новый API принимает запросы прямо в адресе, отдаёт JSON-stat 2, CSV, PX, Excel и HTML, а в бета-версии — Parquet; у Статистики Швеции Parquet пока не включён. Важная оговорка: полноценного сервиса SDMX у PxWeb нет. Старый API умеет отдать таблицу файлом SDMX-ML 2.0, в новом SDMX среди форматов нет вовсе. Для обмена с международными организациями PxWeb обычно дополняют другим инструментом.

- **.Stat Suite** — открытая платформа сообщества SIS-CC, которое ведёт ОЭСР. Она построена вокруг SDMX и собрана из модулей: обозреватель данных (Data Explorer, выпуск 27.7.0 — июль 2026 года), инструмент загрузки и сопровождения данных и служба SDMX. На ней работают OECD Data Explorer, ФАО, МОТ, статистические службы Австралии, Новой Зеландии, Италии, Швейцарии.
- **SDMX-RI** — набор компонентов Евростата под лицензией EUPL. Главный из них — веб-служба, которая берёт данные из базы распространения и отдаёт их по SDMX. Компоненты можно встроить в свою систему по отдельности. Именно эту службу использует .Stat Suite: служба, поставившая .Stat Suite, на деле ставит и SDMX-RI.
- **Fusion Metadata Registry** — реестр структур SDMX: справочников, описаний таблиц и понятий. Его ведёт Банк международных расчётов. Приложение бесплатно по публичной лицензии БМР, сопутствующие инструменты открыты под Apache 2.0, выпуск 12.4.2 вышел 19 сентября 2026 года. На нём работает, например, реестр SDMX Евростата.

Рядом стоят отраслевые и узкие продукты:

- **DHIS2** — система сбора и анализа медицинской статистики Университета Осло, выпуск 2.43 — май 2026 года. На ней министерства здравоохранения десятков стран Африки и Азии собирают отчёты больниц и считают показатели. В реестре 74 установки в 54 странах, в основном в Африке и Азии. Это прежде всего система сбора, публичная витрина у неё — дополнительная возможность: что из данных видно без входа, решает владелец установки.
- **OpenSDG** — статический сайт показателей целей устойчивого развития, выпуск 2.4.0 — август 2025 года. Понятная витрина и полная выгрузка файлами, но не канал обмена. Подробнее — в главе о статических сайтах.
- **PxStat** — платформа статистической службы Ирландии: каждая таблица одновременно набор данных с выгрузкой в JSON-stat, PX, CSV и XLSX. Почти не тиражирована.

Всё остальное — собственные разработки служб, отсюда и 44% самописных витрин.

Ядро и дополнения

У платформ показателей нет экосистемы расширений, как у CKAN. Возможности либо есть в продукте, либо их добавляют соседним продуктом.

Возможность	PxWeb	.Stat Suite	OpenSDG	DHIS2
API	ядро	ядро (SDMX)	нет, вместо него файлы	ядро
SDMX	только файл SDMX-ML 2.0 в старом API	ядро	дополнительная возможность	не проверено
JSON-stat	ядро	не проверено	не проверено	не проверено
Массовая выгрузка	таблица целиком через API в пределах лимита	через API	ядро	через API
Реестр структур	нет	ядро; можно подключить Fusion Metadata Registry	нет	нет
DCAT	нет	не проверено	не проверено	не проверено

Возможность	PxWeb	.Stat Suite	OpenSDG	DHIS2
Отдельный интерфейс	новый интерфейс PxWeb 2 отделён от API	обозреватель данных отделён от службы SDMX	шаблон сайта правится в Git	приложения поверх Web API

Из таблицы и цифр реестра следуют два практических вывода. Выгрузку DCAT реестр нашёл меньше чем у десятка каталогов показателей, на какой бы платформе они ни работали. Чтобы статистика попала на портал открытых данных и в поисковики, описания придётся делать самим или отдавать таблицы в портал открытых данных отдельным каналом, как Центральный банк Бразилии. А сочетание «PxWeb для людей, .Stat Suite или SDMX-RI для обмена» — обычная схема, а не избыточность.

Как выбрать

Выбор начинается не с продукта, а с того, кто забирает данные. Если только люди — хватит витрины. Если программы и международные организации — нужны API, SDMX и массовая выгрузка.

Ситуация	Что взять	Почему и чем платим
Служба публикует таблицы, обмен по SDMX от неё пока не требуют	PxWeb, новые установки — сразу на PxWebApi 2	быстро, API и JSON-stat в ядре; SDMX придётся добавить позже
Нужен обмен с Евростатом, МВФ, ОЭСР	.Stat Suite или SDMX-RI, структуры — в Fusion Metadata Registry	SDMX в ядре; внедрение тяжёлое, без консультантов его ставят редко
Нужна понятная витрина целей устойчивого развития	OpenSDG	статический сайт и полная выгрузка; как единственный канал обмена слаб
Министерство собирает медицинскую статистику	DHIS2	сбор и публикация в одной системе; публичную часть надо настроить отдельно
Город или регион хочет витрину своих показателей	местный поставщик вроде Swing или раздел на региональной платформе	быстро и недорого; данные у поставщика — договоритесь о полной выгрузке заранее
Показателей десятки, обновляются раз в год	без портала: CSV или Parquet со словарём полей на сайте или в Git, плюс передача в национальную службу	дёшево и долговечно; нет конструктора срезов
Нужны графики для руководства	дашборд поверх выгрузки, а не вместо неё	у дашбордов в реестре API почти нет

Своя разработка оправдана, когда служба уже держит команду и хочет витрину, какой нет у готовых продуктов. Образцы — порталы ЕЦБ, Статистики Канады и малайзийский OpenDOSM. Но у всех троих свой сайт стоит поверх открытого API и массовой выгрузки, а не вместо них. Если команда не готова сделать и сопровождать API, своя витрина повторит судьбу 1505 самописных каталогов, из которых программы читают только 8%. Разбор выбора для статистической службы — в разобранном примере, общая логика — в главе «Свои силы, свободное ПО или облачный сервис».

Когда свой каталог не нужен

Ведомству или городу с несколькими десятками показателей отдельный каталог часто не нужен. Можно передать ряды в национальную статистическую службу, выложить их на портал открытых данных таблицами со словарём полей или опубликовать CSV и Parquet в Git-репозитории с описанием показателей. Чтобы такие ряды нашли, у каждого файла нужны методика, единица измерения, коды территорий из общего справочника и дата обновления. Способы публикации без портала — в части «Публикация без портала».

Что отличает удобный каталог показателей

Шесть признаков, по которым видно, что статистика опубликована как данные, а не как сайт с графиками.

1. **Общие справочники.** Территории, отрасли, товары закодированы одними справочниками для всех показателей. Тогда любой ряд читается одной программой. У Евростата это общие справочники SDMX на все наборы.
2. **Массовая выгрузка.** Всю базу можно забрать без кликов по каждому показателю. FAOSTAT отдаёт архив на каждую тематическую базу и машиночитаемый перечень архивов. Евростат публикует опись всех наборов со ссылками на файл каждого в TSV, CSV и SDMX. У МОТ есть отдельная страница массовой выгрузки ILOSTAT.
3. **Общий поиск.** Один поиск находит и показатель в базе, и таблицу в сборнике, и архивную публикацию.
4. **Обычный портал открытых данных как канал.** Показатели выкладываются наборами на стандартной платформе с её API. Так делает Центральный банк Бразилии: его портал открытых данных работает на CKAN.
5. **Несколько форматов и код.** Не только Excel: CSV, SDMX, JSON-stat, Parquet, форматы статистических пакетов, пример запроса на Python или R прямо на странице показателя.
6. **Техническая документация.** Описание API по OpenAPI, примеры запросов, лимиты и лицензия, записанные явно. Документация в виде веб-страниц с примерами, а не PDF-отчёта подрядчика. У PxWebApi 2 Статистики Швеции настройки — лимит в 150 тысяч значений на запрос, 30 запросов за 10 секунд, лицензия CC0 и список форматов — отдаются по отдельному адресу.

Каталог, где есть только конструктор таблиц, не проходит по второму, пятому и шестому пунктам, даже если выглядит современно.

Примеры

Хорошие порталы показателей устроены похоже: одна база и несколько входов для разных людей. Аналитику нужен график, программисту — API, тому, кто строит свою базу, — выгрузка целиком, а автору статьи — постоянная ссылка на ряд.

Национальные службы и центральные банки

- Портал данных Европейского центрального банка обслуживает все эти аудитории сразу. У каждого ряда есть график и постоянный адрес вида `data.ecb.europa.eu/data/datasets/EXR/EXR.D.USD.EUR.SP00.A`, на который можно

сослаться. Для программ есть API по стандарту SDMX 2.1, а каждый набор скачивается целиком. Наборов было 108 и в 2024 году; тогда все вместе они занимали 93 ГБ в CSV и 664 МБ после перевода в Parquet.

- Портал данных БМР — образец работы с SDMX. Есть API, постоянная ссылка на ряд и раздел массовой выгрузки, где каждая тема скачивается одним архивом в CSV или SDMX.
- Статистика Канады превратила сайт службы в каталог данных. В октябре 2026 года через её API доступно 8270 таблиц, из них около 5,1 тысячи действующих, остальные в архиве. Каждая таблица скачивается целиком в CSV или SDMX. К аналитическим текстам приложены таблицы из той же базы.
- Мексиканский INEGI собирает всё по одному обследованию на вкладках одной страницы: документацию с методикой и описанием в стандарте DDI, таблицы, микроданные, открытые данные и инструменты. Есть API банка индикаторов и раздел массовой выгрузки по территории, теме и году.
- Статистика Хорватии справляется с раздробленностью единым каталогом. На странице темы в одном списке собраны оперативные выпуски, сборники, таблицы PC-Axis, базы краткосрочной статистики и слои геостатистики, метаданные которых ведутся в GeoNetwork. Пользователь не обязан знать, в какой системе лежит нужная таблица.
- OpenDOSM — портал статистической службы Малайзии с открытым кодом. Каждый набор лежит в CSV и Parquet, на странице есть пример чтения на Python, для программ есть API. На той же платформе работает национальный портал data.gov.my.
- Статистика Швеции и Статистика Норвегии перевели свои базы на PxWebApi 2: запрос к любой таблице собирается в адресе, ответ по умолчанию в JSON-stat 2.

Платформы и международные базы

- OECD Data Explorer — .Stat Suite.
- Statbank Denmark — PxWeb.
- Браузер данных Евростата — ряды с разрезами, не файлы по годам; API в SDMX и JSON-stat.

Агрегаторы

- DBnomics — свободный агрегатор рядов CEPREMAP: собирает данные статистических служб, центральных банков и международных организаций, отдаёт их через API и клиенты для Python, R, Julia и Stata. Удобен, когда нужны ряды многих стран в одном виде, но это вторичный источник: сверяйте с первоисточником.
- Our World in Data — витрина с собственной платформой, не национальная статистика.

Показатели и микроданные одной службы — два каталога. Почему их не склеивают в один самописный сайт, разобрано в примере статистической службы.

Где ломается

Конструктор вместо данных. Таблица доступна только через конструктор: выгрузка есть, API нет. Методика показателя живёт в PDF, а не рядом с рядом. Число «наборов» раздувают, выкладывая один показатель отдельным файлом на каждый год.

Выгрузка повторяет экран. В российской ЕМИСС (fedstat.ru) в 2022 году показатель открывался в «просмотре по умолчанию» — последние годы и несколько регионов, — и скачивался тоже только этот срез. Чтобы получить ряд целиком, нужно было вручную отметить все разрезы.

Справочники у каждого свои. В той же ЕМИСС на семь тысяч показателей приходилось почти 700 справочников, многие из которых повторяли друг друга. У Евростата в то же время было 283 справочника на 10,5 тысячи наборов. Без общих справочников ряды разных ведомств не склеиваются.

Ряды останавливаются. В июне 2025 года при сплошной выгрузке из ЕМИСС удалось скачать 6905 показателей из 8773 заявленных, и 41,8% из них не обновлялись с 2022 года. У некоторых ведомств ряды остановились гораздо раньше: у Россельхознадзора последнее обновление было в 2014 году. Независимый подсчёт проекта «Если быть точным» (август 2025 года) дал похожую картину: из 7280 доступных показателей актуальных и обновляемых около 3,7 тысячи.

Систему закрывают. В мае 2025 года стало известно, что по приказу Минцифры и Росстата ЕМИСС выводится из эксплуатации до 31 декабря 2025 года. Текста приказа тогда не было ни на сайтах ведомств, ни на портале официального опубликования. Замена — государственная система «Цифровая аналитическая платформа», запуск которой с 2022 года переносили несколько раз; по данным «Если быть точным», переход отложили до января 2027 года. В ноябре 2025 года ЕМИСС ещё работала: по снимку веб-архива — 8875 показателей от 61 ведомства. В 2026 году сайт не открывается из-за рубежа, и проверить его состояние со стороны нельзя. Вывод для пользователя: ряд, который нужен надолго, скачивайте целиком и храните у себя. Как общественный архив сохранил ЕМИСС, рассказано в главе об архивации.

Публикации разбросаны. Статистика разных уровней может не сходиться в одном месте. У территориальных органов Росстата у каждого свой сайт со своим набором сборников, часть которых продаётся по подписке, и единого каталога этих публикаций нет (2024).

Поставщик уходит. Knoema — облачный сервис показателей, на котором в реестре числятся 125 витрин, в основном африканских: статистических служб, центральных банков и платформы Open Data for Africa. В 2020 году компанию купил холдинг Eldridge, в 2023 году она сменила имя на Seek и развернулась к аналитике для розничной торговли. С января 2026 года knoema.com не открывается: последний удачный снимок веб-архива — 1 декабря 2025 года. В октябре 2026 года порталы Open Data for Africa отвечают программам проверкой Cloudflare вместо данных. Для службы, которая держит статистику в облачном сервисе, это главный риск: заранее записать в договор полную выгрузку базы и справочников в открытом формате.

Платформу не обновляют. Тиражируемое ПО помогает, только если его обновлять. Статбанк Армении (statbank.armstat.am) долго работал на старой версии PxWeb на ASP.NET: стандартный API не отвечал, и чтобы забрать данные, приходилось писать отдельный парсер (ноябрь 2025 года). В октябре 2026 года там уже стоит PxWeb выпуска 2025 года, и стандартный API отвечает. Обновление вернуло данные в общие инструменты без новой разработки.

Меняется поколение API. Переход на PxWebApi 2 меняет адреса запросов: они стали стабильнее, но старые скрипты пользователей перестают работать. Статистика Швеции выпустила конвертер старых ссылок в новые. Службе, которая переходит на новое поколение, стоит держать старый API, пока пользователи не переехали, и объявить срок его отключения.

SDMX тяжело внедрять. SDMX нужен для обмена с международными организациями, но внедрять его тяжело. По оценке 2026 года, активно внедряемых серверных продуктов всего три, и без консультантов их ставят редко. Поэтому за пределами национальных служб и международных организаций SDMX почти не встречается, как и более лёгкий JSON-stat. Небольшой службе разумно начать с массовой выгрузки в CSV или Parquet и документированного API, а точку SDMX добавить готовым продуктом, а не писать самой.

Новый портал теряет старое. В 2024 году портал данных МВФ давал наборы для массовой выгрузки. В новом data.imf.org, открытом весной 2025 года, появились каталог наборов, сквозной поиск, дашборды и новый SDMX API. Но в апреле 2025 года скачать всё сразу было нельзя, машиночитаемого перечня наборов не было, разметки Schema.org и DCAT тоже, на отдельный ряд нельзя было дать ссылку, а API требовал регистрации. В октябре 2026 года SDMX API МВФ отвечает без ключа, остальное стоит проверять заново.

Дашборд вместо данных. Power BI, Tableau и R Shiny в реестре — около сотни витрин показателей, и программный доступ есть у единиц. График виден, а ряд за ним достать нельзя. Подробнее — в антипаттерне «Дашборд вместо данных».

Как проверить каталог показателей

Четыре шага на четверть часа.

1. **Возьмите один ряд и попробуйте сослаться на него.** Есть ли у ряда адрес, который откроется у другого человека с тем же срезом?
2. **Скачайте таблицу целиком.** Совпадает ли файл с полной таблицей или только со срезом, показанным на экране?
3. **Найдите API и его документацию.** Нужна ли регистрация, какие лимиты, какая лицензия у данных?
4. **Проверьте справочники.** Скачивается ли список кодов регионов, и одинаковы ли коды в двух разных таблицах?

Типичные адреса у разных платформ:

```
# SDMX REST, ЕЦБ: последнее значение ряда в CSV вместе с атрибутами
curl -s "https://data-api.ecb.europa.eu/service/data/EXR/D.USD.EUR.SP00.A?
lastNObservations=1&format=csvdata"
# SDMX REST: список таблиц (потоков данных) у ОЭСР, МВФ и Евростата
curl -s "https://sdmx.oecd.org/public/rest/dataflow/all"
curl -s "https://api.imf.org/external/sdmx/2.1/dataflow"
curl -s "https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/dataflow/ESTAT/all/latest"
# Евростат: срез таблицы в JSON-stat 2.0
curl -s "https://ec.europa.eu/eurostat/api/dissemination/statistics/1.0/data/demo_pjan?
geo=FR&sex=T&age=TOTAL&lastTimePeriod=1"
# PxWebApi 2: лимиты, форматы и лицензия установки
curl -s "https://statistikdatabasen.scb.se/api/v2/config"
# PxWebApi 2: сколько таблиц в базе
```

```
curl -s "https://data.ssb.no/api/pxwebapi/v2/tables?lang=en&pageSize=1"
```

Ответ по адресу SDMX или PxWebApi — хороший знак. Если API нет, а есть только конструктор, проверьте, нет ли описи для массовой выгрузки, как у Евростата и FAOSTAT. Сводный список — в чек-листах, общие команды — в главе об API и массовой выгрузке.

Чем отличается от соседей

Сосед	Главное отличие
Портал открытых данных	единица публикации — показатель с разрезами, а не файл; одна таблица заменяет сотни файлов по годам и регионам
Каталог микроданных	агрегаты, а не записи о респондентах; доступ обычно открыт без заявок
Геопортал	показатель привязан к коду территории, а не к геометрии; карта строится по справочнику, общему с геопорталом
Каталог метаданных	реестр структур SDMX хранит справочники и описания таблиц, а каталог показателей отдаёт сами значения
Дашборд	у каталога есть поиск, адрес ряда и выгрузка, у дашборда — только картинка

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: число каталогов, владельцы, страны, платформы с долей API, адреса SDMX и DCAT. Запросы — в приложении.
- SDMX, выпуск SDMX 3.1, май 2025 года; JSON-stat; StatDCAT-AP и черновик 3.0.0; RDF Data Cube
- Массовая выгрузка FAOSTAT, опись наборов Евростата, массовая выгрузка ILOSTAT, портал открытых данных ЦБ Бразилии — проверено 12 октября 2026 года
- PxWeb: старая линия и её выпуски, PxWebApi 2 у Статистики Швеции, руководство PxWebApi 2, выпуски PxWebApi и PxWeb 2; настройки API Статистики Швеции и таблицы Статистики Норвегии; ответ старого API Статистики Швеции в формате SDMX-ML 2.0 — проверено 12 октября 2026 года
- .Stat Suite и выпуски обозревателя данных; SDMX-RI и его компоненты под EUPL; образ службы SDMX-RI для .Stat Suite; Fusion Metadata Registry и его лицензия, реестр SDMX Евростата — проверено 12 октября 2026 года
- DHIS2 и его выпуски; OpenSDG; PxStat; DBnomics и его API — проверено 12 октября 2026 года
- Knoema: покупка холдингом Eldridge, 21.12.2020; объединение с Seek, 15.02.2023; снимки knoema.com в веб-архиве: последний удачный, 01.12.2025, далее ошибки 530; ответы knoema.com и порталов opendataforafrica.org — проверено 12 октября 2026 года
- Портал данных ЕЦБ: список наборов с массовой выгрузкой, описание SDMX API; массовая выгрузка БМР, API БМР; API банка индикаторов INEGI, массовая выгрузка INEGI, страница обследования ENIGH 2022; веб-сервисы Статистики Канады; портал статистики Хорватии;

OpenDOSM и его код; Статбанк Армении, SDMX API МВФ — проверено 12 октября 2026 года

- И. Бегтин, Статистика как дата-продукт, август 2024 — ступени публикации статистики и признаки современного портала
- И. Бегтин, Почему невозможно хвалить Росстат, январь 2022, и Российская статистика: немашиночитаемая институциональная фрагментация, декабрь 2024
- «Если быть точным», В 2025 году в России исчезли 140 датасетов и перестали обновляться 425 показателей, 06.08.2025 — состояние ЕМИСС и перенос перехода на ГИС ЦАП; снимок fedstat.ru в веб-архиве, 01.11.2025
- И. Бегтин, Telegram: ЕМИСС: справочники и просмотр по умолчанию, 30.04.2021; формат PX на порталах открытых данных, 08.01.2023; портал данных ЕЦБ, 23.07.2024; объём данных ЕЦБ в CSV и Parquet, 20.08.2024; INEGI, 12.04.2024; Статистика Канады, 20.11.2024; Хорватия, 03.12.2024; OpenDOSM, 24.03.2023; DBnomics, 19.11.2024; тиражируемое ПО статистики, 01.02.2025; данные МВФ, 31.07.2024; новый портал МВФ, 19.04.2025; ЕМИСС выводят из эксплуатации, 12.05.2025; состояние ЕМИСС, 20.06.2025; Статбанк Армении, 27.11.2025; трудность SDMX, 03.09.2026
- Единицы публикации и форматы

Каталоги микроданных

Каталог микроданных описывает переписи и обследования: кого спрашивали, как строили выборку, какие вопросы задавали, что означает каждая переменная. К описанию прилагаются файлы с записями об отдельных людях, домохозяйствах или предприятиях — обезличенные, но всё равно чувствительные. Поэтому описание здесь открыто почти всегда, а файл часто выдают по регистрации, по заявке или только в защищённой среде.

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) таких каталогов 326 — меньше 1% всех каталогов. У половины из них (49%) часть файлов закрыта, и это нормально для типа. А вот доля неработающих сайтов здесь самая высокая среди всех типов: 59 каталогов, около 18%, не открываются.

Главный вывод главы: каталог микроданных ценен описанием, а не файлом. Исследование, которое нельзя найти по переменным и для которого не написано, как получить доступ, для исследователя не существует, даже если файлы лежат на сервере. Хороший каталог открыто показывает всё, кроме самих записей, отдаёт описания программам в стандарте DDI и честно говорит, кто, на каких условиях и в какой срок выдаст данные.

Как устроено описание обследования

В портале открытых данных главное — файл, в каталоге показателей — число с разрезами. Здесь главное — исследование, у которого три уровня описания.

- **Исследование** (study) — одна перепись или одна волна обследования: кто проводил, когда, на какой территории, кого опрашивали (единица наблюдения и генеральная совокупность), как строили выборку, каким способом собирали ответы.
- **Файл данных** — одна таблица записей: домохозяйства, члены домохозяйства, земельные участки. У одного исследования файлов бывает больше сотни.
- **Переменная** — одна колонка файла: имя, подпись, текст вопроса, допустимые значения с подписями, число ответов по каждому значению, пропуски, признак веса.

Рядом лежат **материалы**: вопросники, инструкции интервьюерам, отчёт о выборке, коды для расчёта весов. И **условия доступа**: тип доступа, текст соглашения, как ссылаться на данные.

Как это выглядит вживую, видно на панельном обследовании домохозяйств Нигерии 2023–2024 годов в библиотеке микроданных Всемирного банка. 12 октября 2026 года API каталога отдаёт по нему:

Уровень	Что в описании
Исследование	номер `NGA_2023_GHSP-W5_v02_M`, Национальное бюро статистики Нигерии, единицы наблюдения — домохозяйства, люди, участки и общины; описание выборки; DOI `10.48529/90ec-pv69`
Файлы	108 файлов, у каждого описание: какой раздел анкеты и на каком уровне собраны данные
Переменные	4531 переменная. У переменной `zone` шесть значений — от «1. North Central» до «6. South West» — и число записей по каждому: 851, 824, 870, 835, 839, 848, всего 5067
Доступ	файлы общего пользования: скачать после регистрации и согласия с условиями — ничего не копировать для других, не пытаться узнать респондентов, ссылаться на источник по образцу

Описание целиком скачивается в формате DDI Codebook 2.5, а DOI ведёт на страницу исследования и отдаёт готовую ссылку в BibTeX. По частоте значений исследователь ещё до заявки видит, хватит ли ему наблюдений по нужному региону. Это и есть главная работа каталога микроданных: дать решить, нужны ли данные, не выдавая их.

Что получает свою страницу

В каталоге микроданных свою страницу почти всегда получает исследование, в хороших каталогах — ещё и каждая переменная. Поэтому число «наборов» здесь несопоставимо с порталами открытых данных: одно исследование — это сотня файлов и тысячи переменных. 12 октября 2026 года в библиотеке микроданных Всемирного банка 7145 исследований, в сводном каталоге IHSN — 13 085, в цифровой библиотеке переписей Индии — 40 268 записей. Каталог Maelstrom Research на OBiBa Mica описывает 473 эпидемиологических исследования и около 2,05 миллиона переменных: поиск идёт по переменным, а не по названиям исследований.

Уровни доступа

Уровни доступа — то, что отличает каталог микроданных от остальных типов. Это не недостаток, а способ дать данные тем, кому они нужны, и не раскрыть людей. Удалить имена недостаточно: человека часто выдаёт сочетание возраста, профессии и маленького населённого пункта (подробнее — в статье «Обезличивание» глоссария). Чем подробнее файл, тем строже доступ.

В NADA, самой распространённой платформе типа, семь типов доступа:

Тип доступа	Нужен вход	Что делает пользователь	Одобрение	Откуда данные
Открытые данные	нет	принимает условия	нет	из каталога, в открытых форматах вроде CSV
Прямой доступ	нет	принимает условия	нет	из каталога, часто в форматах SPSS или Stata
Файлы общего пользования	да	заполняет форму запроса	нет	из каталога
Лицензированные данные	да	подаёт заявку с целью исследования	да	из каталога после одобрения
Защищённая среда	нет	скачивает форму заявки	вне сайта	только в защищённом помещении или среде
Внешний репозиторий	нет	переходит по ссылке	—	с сайта владельца данных
Данные недоступны	—	—	—	только описание

Два последних типа важны не меньше остальных. Каталог может быть полезен, даже если сам не хранит ни одного файла: в библиотеке микроданных Всемирного банка 12 октября 2026 года у 67% исследований (4790 из 7145) данные лежат во внешнем репозитории — у статистической службы, в программе DHS или в IPUMS. Файлы общего пользования — у 18%, прямой доступ — у 12%, открытые данные и лицензированный доступ — около 2% каждый.

Самые подробные данные — налоговые, медицинские, связанные регистры — не выдают вообще. С ними работают в **доверенной исследовательской среде**: одобренный исследователь заходит в защищённую систему, считает там и забирает только проверенные результаты. Так устроены SecureLab в UK Data Service и французский CASD. Норвежский microdata.no, который с 2018 года ведут Sikt и Статистика Норвегии, пошёл дальше: сотрудники и студенты одобренных организаций анализируют регистровые данные онлайн без заявки, но видят только результаты расчётов, а не записи. По докладу ОЭСР 2025 года о 16 статистических службах, такая роль «доверенного посредника» становится для служб обычной. Ещё один путь — **синтетические данные**: искусственные записи с теми же статистическими свойствами, которые можно отдать всем.

Для каталога из этого следуют три требования. Описание открыто при любом уровне доступа. Условия записаны до заявки: кто может получить, что нужно приложить, сколько ждать ответа. И на странице видно, какой уровень у каждого файла, а не только у исследования в целом. Как это устроено в разных платформах, разобрано в главе о доступе и ролях.

Кто создаёт

Больше половины каталогов микроданных у центральных органов власти — в основном это национальные статистические службы.

Владелец	Каталогов	Доля
Центральные органы власти	179	55%
Университеты и научные организации	104	32%
Международные организации	20	6%
Общественные организации	14	4%
Регионы	4	1%
Компании и прочие	5	2%

- **Статистические службы** публикуют переписи и обследования домохозяйств, рабочей силы, бюджетов, здоровья. В развивающихся странах каталог часто появлялся при поддержке программ Всемирного банка и сети IHSN, поэтому 59% каталогов служб — на NADA.
- **Университетские архивы социальных данных** хранят опросы общественного мнения, панельные и когортные исследования. Здесь платформы пестрее: NADA только у 21%, рядом OBiBa Mica, Colectica, Dataverse и свои системы.
- **Международные организации** — Всемирный банк, ВОЗ, ФАО, УВКБ ООН, Тихоокеанское сообщество. У них 18 из 20 каталогов на NADA.
- **Консорциумы опросов** — Afrobarometer, Arab Barometer, Latinobarómetro, Asian Barometer — публикуют свои волны на обычных сайтах: на WordPress или собственных.

Каталоги микроданных есть в 139 странах, ещё 27 охватывают весь мир. Ни у одной страны их не больше двух десятков: США — 18, Канада — 12, Германия — 11. Тип устроен так, что у страны обычно один-два каталога статистической службы и несколько университетских архивов.

Почему 326 — нижняя граница

Число каталогов микроданных в реестре сильно занижает мир обследований. Самые крупные архивы социальных данных — UK Data Service, ICPSR, GESIS, финский FSD, шведский SND, Австралийский архив данных, норвежский Sikt — в реестре записаны научными репозиториями или порталами открытых данных: у них DOI, долгое хранение и данные не только из обследований. Европейский каталог CESSDA, который собирает описания национальных архивов, 12 октября 2026 года отдавал около 33 тысяч описаний исследований. Поэтому для поиска обследований стоит смотреть и в главу о научных репозиториях: граница между двумя типами проходит внутри архивов социальных наук.

Что публикуют

- **Переписи населения** — полные или выборочные файлы записей. В Латинской Америке их распространяют через REDATAM, в остальном мире — через NADA и IPUMS.
- **Обследования домохозяйств** — бюджеты, уровень жизни, рабочая сила, здоровье. Многие из них — международные программы с общей анкетой: LSMS Всемирного банка, DHS, MICS ЮНИСЕФ, STEPS ВОЗ.
- **Опросы общественного мнения и панели** — European Social Survey, немецкий SOEP, российский RLMS-HSE, китайский CFPS.
- **Когорты и биобанки** — эпидемиологические исследования, где за людьми наблюдают годами: CanPath и CLSA в Канаде, европейские проекты RECAP Preterm и EUCAN.
- **Микроданные предприятий и банков** — исследовательские центры Бундесбанка, Банка Италии, Банка Португалии.
- **Административные регистры** — налоги, занятость, пособия, связанные по личному коду. Их почти никогда не выдают файлом, только в защищённой среде.
- **Код и данные для перепроверки исследований** — репозиторий воспроизводимых исследований Всемирного банка на той же NADA: 585 записей в октябре 2026 года против 163 в 2024-м.

Чего в каталоге микроданных обычно нет: имён, адресов и точных координат респондентов (их удаляют до публикации), готовых рядов показателей (это каталог показателей), а у закрытых исследований — самих файлов.

Стандарты

DDI (Data Documentation Initiative) — главный стандарт типа. Его ведёт DDI Alliance, а 6 марта 2026 года общая часть стандартов DDI опубликована как ISO/PAS 25955:2026. Это не один формат, а семья:

- **DDI Codebook** — описание одного исследования и его файлов, «электронная книга кодов». Именно его отдают NADA, Dataverse и европейский каталог CESSDA. Версия 2.5 вышла в 2012 году и остаётся основной. Версия 2.6 вышла 15 апреля 2026 года, её поддержка в платформах только начинается.

- **DDI Lifecycle** — описание всего цикла: вопросник, повторное использование вопросов и переменных между волнами, обработка, версии. Версия 3.3 (2020). Нужен лонгитюдным исследованиям и службам, которые ведут банк вопросов. На нём работает Colectica.
- **DDI-CDI** — новая модель для данных любой области, версия 1.0 вышла в 2025 году. Описывает не только обследования, но и связь данных между доменами и происхождение каждого значения. В каталогах пока почти не встречается.

Рядом с DDI работают контролируемые словари — общие списки значений для способа сбора, единицы наблюдения, типа выборки. Их ведут DDI Alliance и CESSDA. Без них поле «способ сбора» в двух архивах заполнено по-разному, и общий поиск по нему не работает.

Обмен описаниями. Сборщики забирают DDI по протоколу OAI-PMH. Так устроен каталог CESSDA: он принимает описания только по OAI-PMH в DDI Codebook 2.5 и по своему профилю метаданных, который проверяет отдельный валидатор. Dataverse отдаёт описания по OAI-PMH в DDI в ядре. У NADA протокола OAI-PMH нет: описания забирают через её собственный API или скачивают в DDI по одному исследованию. Поэтому каталог на NADA нельзя подключить к европейской федерации архивов напрямую, без промежуточной выгрузки.

Для поисковиков нужны не DDI, а разметка schema.org/Dataset и DOI. NADA ставит разметку JSON-LD на страницу исследования в ядре. DOI выдаёт DataCite, как и для научных данных; Всемирный банк к октябрю 2026 года зарегистрировал около 3 тысяч DOI на страницы своей библиотеки микроданных. Подробнее — в главах об экспорте метаданных и об идентификаторах.

Инструменты описания. DDI редко пишут руками. Для обследований есть Metadata Editor Всемирного банка — открытый (MIT) с 2025 года редактор описаний микроданных, документов, таблиц и геоданных; NADA с выпуска 5.7 использует его шаблоны. Для данных, оставшихся в Nesstar Publisher, IHSN выложила на GitHub инструмент выгрузки DDI и данных из его файлов (обновлялся в феврале 2026 года). Коммерческая альтернатива — Colectica Designer.

Платформы

Платформа	Каталогов в реестре	Что это	Не открываются
NADA	158	свободная (MIT) платформа Всемирного банка и IHSN	47
Своя разработка	43	сайты служб и архивов; программный доступ отмечен у 12%	0
REDATAM	38	бесплатная программа ЭКЛАК ООН для переписей, в основном Латинская Америка	6
OBiBa Mica	27	свободный (GPL-3.0) каталог эпидемиологических исследований	3
Nesstar	13	снят с поддержки	2
IPUMS	13	сайты проекта гармонизации Университета Миннесоты	0
Colectica	11	коммерческая платформа на DDI Lifecycle, в основном лонгитюдные исследования	0

Платформа	Каталогов в реестре	Что это	Не открываются
WordPress, Drupal	11 и 7	сайты архивов и опросов	0

Видно, что тип держится на одной свободной платформе. NADA — почти половина каталогов, и почти все неработающие сайты типа — тоже NADA. Это не порок программы, а след того, как её ставили; об этом — в разделе «Где ломается». Самописных каталогов всего 13% — меньше, чем у любого крупного типа: задача узкая, и готовое решение есть. Зато у самописных программный доступ почти нулевой.

Основные платформы

- **NADA** — свободный каталог микроданных, который разрабатывает Всемирный банк вместе с сетью IHSN. Выпуски выходят часто: 5.6 — июль 2026 года, 5.7 — сентябрь, 5.7.1 — 4 октября 2026 года. Уровни доступа, заявки на лицензированные данные, импорт и выгрузка DDI, поиск по переменным — в ядре. С выпуска 5.6 NADA описывает не только обследования, но и показатели с временными рядами, документы, таблицы и геоданные; права можно давать по пользователям и коллекциям, есть встроенная статистика просмотров и скачиваний. Для администраторов есть клиенты API для R и Python (NADAR и PyNADA).
- **REDATAM** — программа отдела народонаселения ЭКЛАК (CELADE), которую развивают с 1985 года. Она хранит переписи в собственном формате и строит по ним таблицы для любой территории, вплоть до квартала. ЭКЛАК распространяет её бесплатно и называет свободной; по данным сайта, она помогла 34 странам. Текущее поколение — RedatamX (с 2024 года): работает в Windows, Linux и macOS и как пакет `redatamx` для R. Предыдущая Redatam7 работает только в 32-битной Windows, последний выпуск — март 2021 года. Онлайн-обработка позволяет считать таблицы прямо на сайте службы без скачивания записей.
- **OBiBa Mica** — свободный каталог эпидемиологических исследований и консорциумов: подробные словари переменных, сети исследований, заявки на доступ. Выпуски выходят регулярно (6.3.7 — сентябрь 2026 года). Работает в паре с хранилищем OBiBa Opal. Данные на таких порталах почти всегда выдают только по заявке.
- **Colectica** — коммерческий набор программ американской компании на DDI Lifecycle: редактор, репозиторий метаданных и портал. На нём работают Статистика Новой Зеландии, статистическая служба Ирландии, британский CLOSER и американские лонгитюдные исследования. В августе 2026 года компания объявила Colectica Repository 8 с сервером MCP, через который ИИ-агент читает описания DDI.
- **IPUMS** — не тиражируемая платформа, а проект: Университет Миннесоты собирает переписи и обследования многих стран и приводит переменные к одному виду, чтобы страны и годы можно было сравнивать. IPUMS International в октябре 2026 года — 104 страны, 656 переписей и обследований, больше миллиарда записей о людях. Данные выдаются бесплатно после регистрации, выборками под запрос.
- **Dataverse** в реестре числится среди научных репозиториев, но для микроданных его выбирают часто: описание в DDI, файлы с ограниченным доступом, DOI и OAI-PMH — в ядре. При загрузке таблиц SPSS или Stata он сам извлекает описания переменных. На

Dataverse переехали канадский ODESI (в репозиторий Borealis) и Австралийский архив данных.

Ядро и дополнения

Возможность	NADA	Dataverse	OBiBa Mica	REDATAM
Уровни доступа к файлам	ядро, семь типов	ядро: открытые и ограниченные файлы, запрос доступа	ядро: описания открыты, данные по заявке	данные на сервере службы, наружу — таблицы
Заявки и их журнал	ядро	ядро	ядро	нет
Описание переменных	ядро, из DDI	ядро, извлекается из файлов SPSS и Stata	ядро, главный объект каталога	ядро, словарь базы
Выгрузка DDI	ядро, DDI Codebook	ядро, DDI Codebook	не проверено	нет, свой формат
OAI-PMH	нет	ядро	не проверено	нет
schema.org	ядро (JSON-LD)	ядро	не проверено	нет
DOI	поле для DOI, выпущенного отдельно в DataCite	ядро	не проверено	нет
Экспорт цитирования	ядро, через DataCite, если есть DOI	ядро	не проверено	нет
API	ядро	ядро	ядро	REST-сервис и пакет для R
Онлайн-расчёты без выдачи записей	нет	нет в ядре	нет	ядро

Из таблицы следуют два вывода. NADA закрывает главное — уровни доступа и описание в DDI, — но в европейскую федерацию архивов и в сборщики научных данных её описания попадают только через промежуточную выгрузку, а DOI придётся регистрировать самим. Dataverse даёт DOI и OAI-PMH, но семи типов доступа и онлайн-расчётов в нём нет. Поэтому статистической службе обычно ближе NADA, а университетскому архиву — Dataverse.

Nesstar: что делать с наследием

Nesstar — набор программ норвежской службы NSD и британского архива данных, на котором в 2000–2010-х годах работали многие архивы и статистические службы. С версией 4.0 в 2015 году его объявили устаревшим, к 2022 году окончательно сняли; Nesstar Publisher распространяется без поддержки. Статистика Канады отключила свой Nesstar в январе 2022 года по соображениям кибербезопасности. Канадский ODESI выбрал для переезда Borealis на Dataverse, и в октябре 2026 года прежний адрес его Nesstar ведёт на новый сайт, а коллекции ODESI видны в Borealis отдельными наборами для OAI-PMH. У GESIS адрес Nesstar-каталога ZACAT ведёт на общий поиск GESIS. UK Data Service ещё держит Nesstar, но в заявлении о доступности от 21 августа 2026 года пишет, что приложение со временем выведут из работы, поэтому исправлять в нём проблемы доступности не станут.

Для нового каталога Nesstar не выбирают. Если каталог ещё на нём, план переезда нужен до того, как сайт перестанет открываться: выгрузить описания в DDI и данные инструментом IHSN, выбрать платформу по задаче (NADA — для каталога с уровнями доступа, Dataverse — для репозитория с DOI) и оставить перенаправления со старых адресов.

Как выбрать

Выбор начинается с двух вопросов: кто будет получать данные и насколько подробны записи.

Ситуация	Что взять	Почему и чем платим
Статистическая служба публикует обследования с разными уровнями доступа	NADA	уровни доступа и DDI в ядре; OAI-PMH и регистрацию DOI придётся добавить самим
Университетский архив, нужны DOI, долгое хранение и обмен с CESSDA	Dataverse или вступление в национальный архив социальных данных	DOI и OAI-PMH в ядре; семи типов доступа нет
Эпидемиологическая когорта или консорциум, данные только по заявке	OBiBa Mica вместе с Opal	подробный поиск по переменным и заявки; данных наружу почти нет, поисковики такие каталоги пропускают
Перепись с таблицами по малым территориям	REDATAM для онлайн-таблиц плюс описание в DDI и файл в открытом формате	быстро даёт таблицы; без открытого файла и DDI данные остаются в формате одной программы
Лонгитюдное исследование с банком вопросов и многими волнами	Colectica или другие инструменты DDI Lifecycle	версии вопросов и переменных между волнами; коммерческая лицензия
Регистры и другие данные, которые нельзя выдать файлом	доверенная исследовательская среда, а каталог — только с описаниями	данные не покидают службу; среда стоит дороже любого каталога
Несколько обследований в год, своей команды нет	передать описания и файлы в национальный архив, IHSN или библиотеку Всемирного банка	дёшево; витрина не своя

Своя разработка в этом типе почти никогда не оправдана: NADA бесплатна, в 2026 году у неё вышли три выпуска, и уровни доступа в ней уже есть. Из 43 самописных каталогов в реестре программный доступ отмечен только у 12%. Если служба держит свою систему, как Статистика Кореи или Австралийское бюро статистики, ей стоит хотя бы отдавать описания исследований в DDI. Как статистической службе разделить показатели и микроданные, разобрано в примере статистической службы, общая логика выбора — в главе «Свои силы, свободное ПО или облачный сервис».

Когда свой каталог не нужен

Исследовательской группе или ведомству с одним-двумя обследованиями свой каталог не нужен. Обследование можно сдать в национальный архив социальных данных или в универсальный репозиторий с ограниченным доступом к файлам, а описание — в каталог IHSN. Если данные можно открыть, хватит файла в CSV или Parquet со словарём переменных и описанием выборки в Git-репозитории или на Zenodo. Чтобы такое обследование нашли, в описании нужны единица наблюдения, территория и период, способ выборки, список переменных и условия доступа. Способы публикации без портала — в части «Публикация без портала».

Что отличает удобный каталог микроданных

1. **Описание открыто всегда.** Исследование видно без входа, даже если файлы закрыты. Тот, кто не знает о данных, не может их запросить.

2. **Поиск по переменным.** У каждой переменной своя страница с текстом вопроса, значениями и частотами. Исследователь находит нужный вопрос, а не перебирает названия обследований.
3. **Условия записаны заранее.** Тип доступа виден у каждого файла, условия и срок ответа на заявку — до заявки.
4. **Описания для программ.** Выгрузка DDI, API каталога, а для поисковиков — разметка schema.org и постоянный идентификатор.
5. **Как ссылаться.** У исследования есть DOI или постоянный адрес и готовая ссылка; для повторных выпусков — версии.
6. **Открытые форматы.** Открытые файлы лежат в CSV или Parquet, а не только в форматах SPSS и Stata, которые требуют конкретной программы.

Примеры

Международные каталоги

- Библиотека микроданных Всемирного банка — NADA, 7145 исследований из многих стран; две трети ведут на сайты владельцев данных, у части есть DOI.
- Сводный каталог IHSN — NADA, 13 085 описаний обследований, в основном развивающихся стран.
- Каталог микроданных ВОЗ по неинфекционным болезням — NADA, 937 обследований STEPS, о табаке и о здоровье школьников. У ВОЗ это один из нескольких каталогов разных типов, как и у Тихоокеанского сообщества с его Pacific Data Hub.
- IPUMS International — гармонизированные переписи 104 стран.

Национальные каталоги

- Цифровая библиотека переписей Индии — NADA, 40 268 записей, переписи с 1901 года с подробностью до округов. В 2023 году записей было больше 29 тысяч.
- Каталог микроданных Армстата — NADA, 26 обследований.
- Национальные статистические службы Боливии (121 исследование) и Танзании (62) — NADA с файлами разных уровней доступа.
- INEGI — микроданные на вкладке каждого обследования рядом с таблицами и описанием в DDI; разобран в главе о каталогах показателей.

Архивы социальных данных

- UK Data Service — открытые, охраняемые и контролируемые данные; самые подробные — только в среде SecureLab.
- Каталог CESSDA — общий поиск по европейским архивам: около 33 тысяч описаний, собранных по OAI-PMH в DDI.
- ODESI — канадский поиск по социальным обследованиям, данные которого с Nesstar переехали в Borealis на Dataverse.

Медицина и защищённый доступ

- Maelstrom Research — OBiBa Mica, 473 эпидемиологических исследования и около 2,05 миллиона описанных переменных.
- microdata.no — норвежские регистровые данные с 1964 года: анализ онлайн без заявки, наружу — только результаты.

Россия

- RLMS-HSE — Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ, панельное обследование с 1994 года; данные после регистрации.
- База данных ФОМ и банк социологических данных Института социологии ФНИСЦ РАН — опросы общественного мнения на своих сайтах.
- Микроданные выборочных обследований Росстата публикуются на отдельных страницах сайта службы и в системе obdx.gks.ru. В 2023 году данные в ней последний раз обновлялись в 2021 году, а программа не менялась около девяти лет. В октябре 2026 года оба адреса из-за рубежа не открываются, проверить их состояние со стороны нельзя.

Где ломается

Установку бросают после проекта. Из 59 неработающих каталогов микроданных 47 — на NADA. Почти все это старые каталоги статистических служб Африки, Азии и Латинской Америки: на IP-адресах вместо доменов, на поддоменах подрядчиков, без HTTPS. Среди каталогов без HTTPS не работают 38 из 58, среди каталогов с HTTPS — 21 из 268. Каталог ставили в рамках программы поддержки, а когда программа закончилась, сопровождать его было некому. У 18 из 38 стран с брошенной NADA другого действующего каталога микроданных в реестре нет. Бывает и так, что сайт открывается, а каталог пуст: NADA статистической службы Мадагаскара 12 октября 2026 года отвечает, но в ней ноль исследований. Урок: каталог микроданных нужно ставить на домен службы и закладывать сопровождение в бюджет, а не только установку.

Продукт снимают. Nesstar — пример закрытия самой платформы. В реестре это видно слабее, чем кажется: из 13 записей на Nesstar неработающими отмечены две, но ещё два адреса — у GESIS и ODESI — на деле ведут на новые системы. Главный риск не в том, что сайт пропадёт, а в том, что описания и файлы остались в формате, который больше никто не поддерживает.

Формат одной программы. Через REDATAM с 1990-х годов распространяют переписи стран Латинской Америки, но база хранится в собственном формате без опубликованного описания: таблицы строят в самой программе или в её онлайн-обработке. Чтобы работать с записями в обычных инструментах, исследователи сделали открытый конвертер Open Redatam (Apache 2.0): он читает словарь базы и выгружает таблицы в CSV, есть пакеты для R и Python. Есть и пакет `redatamx` для R — обёртка над самим REDATAM, права на ядро у CELADE. Оба пути работают, но без них записи недоступны обычным программам.

Каталог закрыт для программ. 12 октября 2026 года каталоги микроданных ФАО и Тихоокеанского сообщества отвечают программам проверкой Cloudflare, а каталог УВКБ ООН — ошибкой 403, причём закрыт и API. Люди в браузере данные видят, а сборщики, поисковики данных и веб-архивы — нет. Для каталога, ценность которого в описании, это особенно обидно: исследование нельзя найти там, где его ищут.

Описания есть, данных нет. Порталы на OBiBa Mica подробно описывают переменные, а данные выдают только по заявке. Поэтому поисковики данных их часто пропускают: файлов нет, есть только описания. Для каталога микроданных это нормально, но описание должно быть полным, иначе исследование не найдут и не запросят.

Описание спрятано или пусто. Исследование описано только названием и аннотацией, без переменных и выборки, — по нему нельзя решить, нужны ли данные. Хуже, когда описание видно только после входа: тогда исследование нельзя найти вообще.

Доступ формально есть. Заявка принимается, но срок ответа не назван, критериев одобрения нет, и ответа нет месяцами. Доступ формально есть и практически нет. Срок ответа и порядок обжалования должны быть написаны на странице исследования.

Ссылка ведёт наружу. Когда каталог в основном указывает на внешние репозитории, как у Всемирного банка, его качество зависит от чужих сайтов. Если владелец данных переехал, запись в каталоге остаётся, а ссылка ведёт в никуда. Такие ссылки нужно регулярно проверять, как в любом каталоге со ссылками вместо файлов — см. главу о внутреннем хранилище.

Как проверить каталог микроданных

Пять шагов на полчаса.

1. **Найдите исследование без входа.** Видны ли выборка, единица наблюдения и период? Есть ли список переменных?
2. **Откройте одну переменную.** Есть ли текст вопроса, значения с подписями и число ответов?
3. **Прочитайте условия доступа.** Указан ли тип доступа у файлов, что нужно для заявки и сколько ждать?
4. **Выгрузите описание программой.** Отдаёт ли каталог DDI, есть ли API или OAI-PMH, открывается ли он без проверки «я не робот»?
5. **Проверьте ссылку.** Есть ли DOI или постоянный адрес, ведёт ли DOI на страницу исследования?

Типичные адреса у разных платформ:

```
# NADA: сколько исследований и сколько из них с лицензированным доступом
curl -s "https://microdata.worldbank.org/index.php/api/catalog/search?ps=1"
curl -s "https://microdata.worldbank.org/index.php/api/catalog/search?ps=1&dtype=licensed"
# NADA: описание исследования, его файлы и одна переменная
curl -s "https://microdata.worldbank.org/index.php/api/catalog/NGA_2023_GHSP-W5_v02_M"
curl -s "https://microdata.worldbank.org/index.php/api/catalog/NGA_2023_GHSP-
W5_v02_M/data_files"
curl -s "https://microdata.worldbank.org/index.php/api/catalog/NGA_2023_GHSP-
W5_v02_M/variables/V2"
# NADA: описание целиком в DDI Codebook (в адресе — числовой номер исследования)
curl -s "https://microdata.worldbank.org/index.php/metadata/export/6410/ddi"
# DOI: куда ведёт и готовая ссылка в BibTeX
curl -sL -H "Accept: application/x-bibtex" "https://doi.org/10.48529/90ec-pv69"
# OBiBa Mica: число исследований, наборов и переменных
curl -s "https://www.maelstrom-research.org/ws/studies/_rql?query=study(limit(0,1))"
# Dataverse и CESSDA: форматы описаний по OAI-PMH, среди них DDI
```

```
curl -s "https://borealisdata.ca/oai?verb=ListMetadataFormats"
curl -s "https://datacatalogue.cessda.eu/oai-pmh/v0/oai?verb=ListMetadataFormats"
```

Все команды проверены 12 октября 2026 года. Ответ JSON по адресу `api/catalog/search` — признак NADA, ответ `_rql` — признак OBiBa Mica. Страница «Just a moment…» вместо ответа — признак того, что каталог закрыт для программ. Сводный список — в чек-листах, общие команды — в главе об API и массовой выгрузке.

Чем отличается от соседей

Сосед	Главное отличие
Каталог показателей	записи о людях и домохозяйствах, а не агрегаты; описание — исследование и переменные, а не ряд с разрезами
Научный репозиторий	уровни доступа и обезличивание обязательны, DOI — желателен; описание в DDI, а не только в DataCite. Архивы социальных наук стоят на стыке двух типов
Портал открытых данных	закрытый доступ к файлам — норма, а не признак беды; описание гораздо подробнее
Каталог метаданных	тоже показывает описания закрытых данных, но не ведёт заявки и не выдаёт файлы

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: число каталогов, владельцы, страны, платформы, неработающие каталоги, каталоги без HTTPS. Запросы — в приложении.
- DDI: DDI Codebook, DDI Lifecycle, DDI-CDI, ISO/PAS 25955:2026, контролируемые словари — проверено 12 октября 2026 года
- CESSDA: OAI-PMH каталога, требования к поставщикам описаний, профили метаданных; число описаний — по OAI-PMH и API каталога, 12 октября 2026 года
- NADA: код и выпуски, типы доступа, документация; разметка schema.org, поле DOI и экспорт цитирования через DataCite — по исходному коду NADA 5.7.1; Metadata Editor, выгрузка из Nesstar Publisher
- Библиотека микроданных Всемирного банка: число исследований и типов доступа, описание панельного обследования Нигерии 2023–2024 и его DOI — по API каталога и DataCite, 12 октября 2026 года; каталог IHSN, цифровая библиотека переписей Индии, Армстат, ВОЗ, репозиторий воспроизводимых исследований Всемирного банка — по API, 12 октября 2026 года
- REDATAM: сайт, вопросы и ответы, пакет redatamx; Open Redatam и пакет redatam — проверено 12 октября 2026 года; M. Vargas Sepúlveda, L. Barkai, The REDATAM format and its challenges for data access and information creation in public policy, Data & Policy, 2025
- OBiBa Mica и её выпуски; Maelstrom Research — по API, 12 октября 2026 года; Colectica Portal, Colectica Repository 8 и MCP, 12.08.2026; IPUMS International

- Nesstar: история продукта, карточка Nesstar Publisher в каталоге CESSDA, отключение Nesstar Статистики Канады, январь 2022; перенос ODESI на Borealis; заявление UK Data Service о доступности Nesstar, 21.08.2026; ответы адресов Nesstar у ODESI и GESIS — проверено 12 октября 2026 года
- Доступ: microdata.no, CASD; OECD, National statistical offices as emerging trusted intermediaries in data governance, OECD Digital Economy Papers, No. 378, 10.09.2025
- Ответы каталогов ФАО, Тихоокеанского сообщества и УВКБ ООН программам, выборочная проверка неработающих каталогов на NADA, каталог статистической службы Мадагаскара — 12 октября 2026 года
- И. Бегтин, Telegram: каталоги микроданных и Росстат, 22.03.2023; цифровая библиотека переписей Индии, 08.04.2023; REDATAM, 22.08.2023; репозиторий воспроизводимых исследований Всемирного банка, 28.10.2024; порталы на OBiBa Mica, 19.12.2024; Redatam и Open Redatam, 04.06.2025; каталоги данных ВОЗ, 15.09.2026
- Доступ, роли и процессы

ML-каталоги

ML-каталог публикует наборы данных для обучения и проверки моделей машинного обучения, а часто и сами модели. Набор здесь нужен не для того, чтобы его прочитал человек или на него сослались в статье, а чтобы его загрузил в код другой исследователь и сравнил свою модель с чужими на тех же данных.

Главный практический вывод главы: почти всем, кто публикует такой набор, свой каталог не нужен. Наборы ищут на нескольких больших площадках, и туда же смотрят конференции, библиотеки и поисковики. Свой каталог оправдан, когда данные нельзя отдать на чужую площадку, а копию выпуска для долгого хранения кладут в научный репозиторий.

Масштаб это подтверждает. 12 октября 2026 года на Hugging Face было 1 085 895 наборов и 3 143 746 моделей, на OpenML — 6434 действующих набора, в архиве UCI — 689. В реестре dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) ML-каталогов всего 100. Реестр считает сайты, а не наборы: Hugging Face и Kaggle в нём — по одной записи на огромные площадки, рядом с сайтами одного бенчмарка из нескольких наборов. По числу строк реестра о размере этого мира судить нельзя.

Что делает каталог ML-каталогом

От портала открытых данных и научного репозитория ML-каталог отличают пять вещей.

- **Набор разбит на части.** Обучающая часть (train) нужна, чтобы обучить модель, проверочная (validation) — чтобы подобрать настройки, тестовая (test) — чтобы честно измерить результат. Разбиение фиксируют заранее, иначе результаты разных авторов несравнимы.
- **Набор загружают из кода.** Вместо ссылки на архив на странице набора — строка кода, которая загружает его библиотекой: `datasets` у Hugging Face, `kagglehub` у Kaggle, `openml` у OpenML.
- **У набора есть карточка.** Это описание состава, способа сбора, известных перекосов и ограничений: на каких данных модель будет работать хуже и для чего набор не годится.
- **Лицензия отвечает на вопрос об обучении.** Можно ли обучать на наборе коммерческую модель, можно ли выкладывать производные наборы, нужно ли указывать источник.
- **Рядом модели и результаты.** Какие модели обучены на наборе, какие результаты они показали на тестовой части. Набор, по которому ведут общую таблицу результатов, называют бенчмарком.

Если ничего из этого нет, перед вами обычный каталог, где среди прочего лежат наборы, пригодные для машинного обучения. Так бывает с порталами открытых данных: на них много таблиц, на которых учат модели, но разбиения на части и карточки с перекосами там нет.

Виды ML-каталогов

Сто записей реестра пёстрые: это восемь разных видов сайтов. Разметка ниже сделана для этой книги по названиям и адресам записей, в самом реестре такого поля нет.

Вид	Что это	Примеры	Записей
Коллекция организации	университет, лаборатория или компания выкладывает свои наборы	архив UCI, OpenML, Monash Forecasting, Yandex Research, наборы ВМК МГУ	23
Хаб	общая площадка, куда любой загружает наборы и модели	Hugging Face, Kaggle, ModelScope, OpenDataLab, Roboflow Universe	15
Площадка соревнований и бенчмарков	задача, данные, скрытая тестовая часть и таблица результатов	Codabench, EvalAI, Grand Challenge, Zindi, DrivenData, Tianchi	14
Сайт одного бенчмарка	страница проекта с несколькими наборами и таблицей результатов	Open Graph Benchmark, Therapeutics Data Commons, DataComp	14
Государственная платформа данных для ИИ	наборы, которые государство собирает или финансирует для своих разработчиков	AI Hub (Корея), AIKosh (Индия), NAIRR (США), AI-on-Demand (ЕС), реестр ЦИТ (Россия)	13
Каталог библиотеки	список наборов и моделей, которые загружает конкретная библиотека	TensorFlow Datasets, Keras, PyTorch Hub, NVIDIA NGC	9
Список и путеводитель	подборка ссылок на наборы, в основном на других хабах	каталоги наборов Hugging Face по языкам, Open Corpus Registry, Post-Training Data Atlas	9
Продавец данных	компания продаёт готовые и заказные обучающие наборы	Datatang, GTS AI	3

От вида зависит, что проверять. У хаба — карточку и лицензию конкретного набора, а не площадки в целом. У площадки соревнований — условия использования данных после окончания соревнования. У каталога библиотеки — откуда библиотека на самом деле скачивает файлы. У государственной платформы — кто вообще может получить данные. Продавцы данных ближе к маркетплейсам, о них — в главе о других типах.

Кто создаёт

ML-каталоги ведут прежде всего университеты и компании. Государств здесь меньше, чем в других типах, но их доля растёт.

Владелец	Записей
Университеты и научные организации	37
Компании	31
Сообщества	16
Центральные органы власти	8
Общественные организации	5
Регионы	2
Международные организации	1

По странам треть записей — международные (34). Дальше США (19), Китай (12), Россия (6), Япония и Корея (по 5), Германия и Индия (по 4). У Китая своя экосистема хабов и площадок соревнований: ModelScope компании Alibaba, OpenDataLab и OpenXLab Шанхайской лаборатории ИИ, платформа BAAI, Tianchi, Baidu AI Studio, HeyWhale и WiseModel. Российские записи — наборы Яндекса и лаборатории компьютерной графики ВМК МГУ, московские медицинские наборы MosMedData, раздел наборов портала ИИ Москвы, Medical Data Hub и реестр ЦИТ Минпромторга.

Самописных сайтов среди ML-каталогов 78, ещё 9 сделаны генератором статических сайтов Jekyll на GitHub Pages — это больше половины сайтов отдельных бенчмарков. Программный интерфейс отмечен только у 23 записей. Готовые платформы встречаются в основном у соревнований: CodaLab, Codabench, EvalAI и Grand Challenge.

Что публикуют

- **Обучающие и тестовые наборы**: тексты, изображения, речь, видео, таблицы, графы, молекулы. Большие текстовые наборы для языковых моделей занимают терабайты, табличный набор из архива UCI — килобайты.
- **Модели** — файлы весов обученной сети с карточкой модели: на чём обучена, для чего годится, где ошибается. На Hugging Face моделей втрое больше, чем наборов.
- **Задачи и результаты** — условия соревнования, метрика, по которой сравнивают модели, и таблица результатов.
- **Производные наборы** — подвыборки, переводы, переразметка чужих наборов и синтетические данные, которые сгенерировала другая модель. Сейчас это, пожалуй, самая быстрорастущая часть, и именно у неё хуже всего с происхождением.

Отдельный случай — каталог библиотеки. TensorFlow Datasets или `torchvision` хранят не данные, а код загрузчика: адрес исходного файла, контрольную сумму и правила разбиения. Файл по-прежнему лежит у автора набора или на зеркале, и когда исходный сайт пропадает, ломается загрузка у всех пользователей библиотеки. Пример — в разделе «Где ломается».

Стандарты

Карточка набора. Это текст для человека: что в наборе, как и зачем его собрали, кто размечал, какие известны перекосы, для чего набор использовать не стоит. Идею предложили в статье «Datasheets for Datasets» (Гебру и соавторы, 2018 год), похожие шаблоны — Data Cards у Google и карточки моделей (2019 год). На Hugging Face карточка — это файл `README.md` в репозитории набора. Вверху файла раздел метаданных: лицензия, языки, задачи, размер, а также какие файлы относятся к какой части набора. По этим полям работают фильтры поиска.

Croissant. Машиночитаемое описание ML-набора, которое ведёт консорциум MLCommons. Оно построено поверх schema.org/Dataset и добавляет то, что нужно программе для загрузки: какие файлы входят в набор, какие в них поля и типы, как набор делится на части, где метки. Файл Croissant — это одновременно обычная разметка schema.org, поэтому его читают и поисковики: Google Dataset Search по нему индексирует наборы и умеет показывать только наборы с Croissant.

- Версия 1.0 вышла в 2024 году, версия 1.1 — 29 января 2026 года. В 1.1 добавлены происхождение данных по словарю W3C PROV-O (из чего и каким процессом получен набор или отдельное поле), условия использования через словари DUO и ODRL (например, «только для исследований, не для коммерции»), правила версий набора и контрольные суммы файлов, пометка для наборов, которые постоянно пополняются.
- Расширение RAI (Responsible AI, «ответственный ИИ», версия 1.0 от марта 2024 года) описывает ограничения набора, перекосы, чувствительные сведения, назначение, синтетическое ли это данные и из каких наборов он сделан.
- GeoCroissant 1.0 (январь 2026 года) — расширение для геоданных: система координат, охват, снимки.

Реализации отстают от спецификации. 12 октября 2026 года Hugging Face, Kaggle, OpenML, Harvard Dataverse и расширение ckanext-dcat отдавали описания с пометкой версии 1.0. Если вам нужны поля о происхождении и условиях использования из 1.1, заполните их сами и проверьте, что их не теряет площадка.

Требования конференций. Главная научная конференция по машинному обучению NeurIPS требует файл Croissant в своём треке о данных. В 2026 году трек называется Evaluations & Datasets, и файл нужен у каждой подачи, которая вводит новый набор. К обязательным полям Croissant в том же году добавлены минимальные поля RAI. Предпочтительные площадки по правилам трека — Kaggle, OpenML, Hugging Face и Dataverse: они создают основные поля Croissant сами, а поля RAI автор дописывает. Если набор лежит на своём сайте, файл Croissant автор делает целиком сам. Для каталога это значит, что поддержка Croissant — уже не украшение, а условие, при котором статью с новым набором примут в этот трек.

Форматы. Таблицы на хабах всё чаще хранят в Parquet: в нём есть типы колонок, а программы читают его по частям, не скачивая целиком. Hugging Face сам переводит наборы в Parquet, но не больше 5 ГБ: если набор больше, переведённая копия будет неполной. Описание Croissant Hugging Face создаёт только для наборов, которые удалось перевести, поэтому архив картинок в ZIP останется без него. OpenML исторически хранит таблицы в ARFF, формате программы Weka, и рядом отдаёт Parquet. О форматах в целом — в главе о форматах и единицах публикации.

Лицензии. Кроме обычных Creative Commons и свободных лицензий программ здесь встречаются лицензии с ограничениями использования, например OpenRAIL (запреты на определённые применения модели), пометки «только для исследований» и правила соревнований, которые часто разрешают использовать данные только в рамках соревнования. Лицензия набора не перекрывает права на исходный материал: набор картинок под CC BY не делает свободными сами картинки, если их собрали из интернета.

Раскрытие обучающих данных. С 2 августа 2025 года в ЕС действует обязанность разработчиков моделей общего назначения по закону об ИИ (AI Act, статья 53) публиковать сводку данных, на которых обучена модель. Шаблон Еврокомиссия опубликовала 24 июля 2025 года, использовать его обязательно; для моделей, выпущенных до 2 августа 2025 года, срок — 2 августа 2027 года. В сводке перечисляют основные наборы и другие источники. Чтобы такую сводку можно было составить и проверить, у набора нужны постоянное название, адрес, версия и лицензия — то есть ровно то, что даёт хорошая карточка в каталоге.

DOI. В ML-каталогах это не правило, а исключение. Hugging Face выдаёт DOI через DataCite по кнопке в настройках набора; после этого набор нельзя удалить, переименовать или скрыть без обращения в поддержку. На 12 октября 2026 года у Hugging Face в DataCite 5433 DOI наборов — около половины процента от их числа. Kaggle выдал 66 208 DOI наборов, архив UCI — 693. О том, как DOI устроены, — в главе об идентификаторах.

Платформы

Здесь два разных выбора: опубликовать набор на чужой площадке или поставить ML-каталог у себя.

Чужие площадки

Площадка	Что это	Загрузка в код	Croissant	DOI	Ограничения
Hugging Face Hub	облачный хаб наборов и моделей	библиотека `datasets`, API, адреса `hf://` в DuckDB, pandas и Polars	создаётся сам для наборов, переведённых в Parquet	по кнопке, через DataCite	доступ по запросу включает автор; бесплатное место для открытых наборов — «по возможности», для закрытых 100 ГБ
Kaggle	облачная площадка наборов, моделей и соревнований	`kagglehub`, Kaggle API	на странице набора и по адресу `/croissant/download`	выдаёт, префикс 10.34740	до 200 ГБ на набор; файл скачивают после входа
OpenML	свободная платформа наборов, задач и экспериментов (BSD-3)	библиотека `openml`, API	по адресу `/croissant/dataset/` и номер набора	в документации не найден	в основном таблицы; задача фиксирует разбиение и метрику
Codabench	свободная платформа соревнований и бенчмарков (Apache-2.0)	—	нет, это не каталог наборов	нет	тестовую часть можно скрыть от участников
Zenodo	общий научный репозиторий	API	нет	у каждой версии	до 50 ГБ на запись бесплатно

Codabench — преемник CodaLab Competitions. CodaLab перестал принимать новые соревнования в сентябре 2025 года и закрыл приём решений 31 декабря 2025 года. Задачи переносятся на Codabench без изменений, а таблицы результатов — нет: решения приходится присылать заново.

Площадки дают то, чего у своего каталога долго не будет: аудиторию, загрузку из кода, предпросмотр таблицы, Croissant и учёт скачиваний. Цена — зависимость от одной компании. О том, чем это рискует и как подстраховаться, — в главе об облачных хранилищах и хабах.

Свой каталог: что в ядре, а что нет

Свою установку обычно делают на платформе научного репозитория или портала открытых данных и добавляют Croissant. Специальной свободной платформы ML-каталога, которую ставят у себя так же просто, как CKAN или Dataverse, нет. OpenML можно развернуть у себя, но это сложная система для экспериментов, а не только каталог.

Возможность	Hugging Face	Kaggle	OpenML	Dataverse	CKAN
Croissant	ядро	ядро	ядро	ядро с версии 6.10, раньше внешний экспортёр	расширение ckanext-dcat, плагин `croissant`
Разбиение на части	ядро: части описаны в карточке	нет: только файлы набора	ядро: задача задаёт разбиение	нет	нет
Загрузка из кода одной строкой	ядро	ядро	ядро	нет: API и клиент pyDataverse	нет: API и прямые ссылки на файлы
Доступ по запросу	ядро: автор включает запрос доступа	нет: только закрытый набор целиком	не найден	ядро: закрытые файлы с запросом доступа	только закрытый набор целиком
DOI	ядро	ядро	не найден	ядро	расширение ckanext-doi

Подробности по своим установкам:

- **CKAN.** Плагин `croissant` в расширении ckanext-dcat появился в версии 2.3.0 (февраль 2025 года), последний выпуск расширения — 2.4.4 от 11 августа 2026 года. Описание встраивается в страницу набора и отдаётся по адресу `/dataset/<id>/croissant.jsonld`. Поля таблицы (`RecordSet` в терминах Croissant) попадают в описание, только если файл загружен в хранилище таблиц CKAN DataStore. Иначе программа узнает о файле, но не о его колонках. Про само расширение и его совместимость с версиями CKAN — в главе про CKAN.
- **Dataverse.** С версии 6.10 Croissant встроен в ядро: в страницу набора встраивается краткая разметка без списка файлов, а полное описание отдаёт выгрузка. Раньше это был внешний экспортёр сообщества GDCC (последний выпуск — 0.1.6), который работал с Dataverse 6.0 и новее и встраивал описание в страницу с версии 6.2; теперь его репозиторий помечен как устаревший. На установках старше 6.10 Croissant по-прежнему ставят этим экспортёром. Harvard Dataverse 12 октября 2026 года отдавал Croissant по запросу к API.
- **GeoNetwork** встраивает в страницу записи разметку schema.org. Поля Croissant для схемы ISO 19115-3 добавлены в ветку будущего выпуска 4.4.13, в выпуске 4.4.12 их нет. Это пригодится, если наборы для обучения — снимки и слои.
- **InvenioRDM и Zenodo** Croissant не отдают: в коде платформы есть разметка schema.org, Data Package и RO-Crate, но не Croissant.
- **Статический сайт.** Девять сайтов бенчмарков в реестре сделаны на Jekyll и лежат на GitHub Pages. Для коллекции из нескольких наборов этого достаточно, если рядом с каждым набором лежит файл Croissant, проверенный валидатором, и сами файлы — на хабе или в репозитории с DOI. Как сделать такой сайт — в главе о статических сайтах.

Государственные платформы данных для ИИ

Государства всё чаще строят свои ML-каталоги: собирают наборы на национальных языках, по медицине, транспорту и промышленности и отдают их своим разработчикам. Устроены они иначе, чем хабы, и ближе к каталогам микроданных: главное в них — кто имеет право получить данные.

- **AI Hub, Корея.** Ведёт Национальное агентство информационного общества (NIA) по заказу Министерства науки и ИКТ. На страницах наборов прямо сказано, что заявку на данные могут подать только граждане Кореи. Иностранной организации нужен отдельный договор с исполнителем и NIA, вывоз данных за границу — тоже по договору. Медицинские данные открывают в «безопасной зоне»: на сервере NIA без скачивания, а обученную модель выдают после проверки.
- **AIKosh, Индия.** В 2026 году — больше 12 тысяч наборов и моделей. У набора есть оценка качества и условия использования, скачивание — после регистрации, лицензия обычно запрещает коммерческое использование. Рядом — платформа языковых наборов Bhashini Vatika.
- **США и ЕС.** Пилот национального ресурса для исследований ИИ NAIRR даёт перечень наборов, моделей и вычислительных ресурсов партнёров. Европейский каталог AI-on-Demand собирает наборы, модели и инструменты проектов ЕС.
- **Россия.** Реестр отраслевых моделей машинного обучения и наборов данных ведёт ФГАУ «ЦИТ» при Минпромторге. В декабре 2025 года в нём было 42 набора и 7 моделей, 12 октября 2026 года — 141 набор. Среди них есть копии чужих наборов с Hugging Face и Kaggle со ссылкой на источник в описании, а лицензия в карточке записана словами «Свободная лицензия» без названия. Москва публикует раздел наборов на портале ИИ города и наборы медицинских снимков MosMedData.

Вывод для такой платформы: если данные закрыты, открытой должна быть хотя бы карточка — с лицензией, происхождением и порядком доступа, в Croissant и schema.org. Тогда набор находят снаружи, а исследователь понимает, стоит ли подавать заявку. Как устроен доступ по заявке и защищённая среда, разобрано в главе о каталогах микроданных и в главе о доступе и рабочих процессах. Тренд в целом — в главе об ИИ и пространствах данных.

Примеры

Хабы

- Hugging Face Datasets — наборы и модели в одном месте, загрузка из кода, Croissant и DOI.
- Kaggle Datasets — наборы, блокноты и соревнования.
- ModelScope — китайский хаб наборов и моделей компании Alibaba.

Коллекции и архивы

- UC Irvine Machine Learning Repository — один из старейших архивов ML-наборов, с API и DOI у наборов.
- OpenML — наборы вместе с задачами, разбиениями и результатами экспериментов.

Соревнования и бенчмарки

- Codabench — соревнования и бенчмарки со скрытой тестовой частью.
- Open Graph Benchmark — сайт одного бенчмарка на GitHub Pages: наборы, загрузчик и таблица результатов.

Каталоги библиотек

- TensorFlow Datasets — список наборов, которые загружает библиотека TensorFlow.

Государственные платформы

- AI Hub — Корея.
- AIKosh — Индия.
- Реестр отраслевых моделей и наборов данных ЦИТ — Россия.

Где ломается

Лицензия неясна или неверна. Исследование Data Provenance Initiative (Лонгпре и соавторы, 2023 год, журнальная версия в Nature Machine Intelligence, 2024) проследило происхождение больше 1800 текстовых наборов для дообучения языковых моделей. У популярных наборов на GitHub и Hugging Face больше 70% лицензий не были указаны вовсе. Там, где лицензия была, у 66% наборов на Hugging Face она относилась к другой категории, чем лицензия автора исходных данных, и обычно была свободнее. Набор перекладывают с площадки на площадку, и ограничения теряются по дороге.

Карточка пустая. По исследованию карточек Hugging Face (ICLR 2024, данные на март 2023 года) из 24 065 репозиториев наборов непустая карточка была только у 30,9%. Зато на них приходилось 95% скачиваний. Все рекомендованные разделы заполнили у 86% из сотни самых скачиваемых наборов и только у 7,9% наборов без скачиваний. Раздел о том, чем опасно использование данных, у самых популярных наборов занимал всего 2,1% текста. Хаб показывает такие наборы в поиске наравне с хорошо описанными.

Набор отозвали, а копии живут. 19 декабря 2023 года Стэнфордская обсерватория интернета нашла в наборе LAION-5B ссылки на изображения сексуального насилия над детьми. LAION в тот же день снял набор и его производные. 30 августа 2024 года вышла исправленная версия Re-LAION-5B, из неё удалено 2236 ссылок. Модели, обученные на старой версии, и копии старого набора на других площадках остались. Хороший каталог в таком случае не удаляет страницу набора, а оставляет её с пометкой «отозван», причиной и ссылкой на замену. Тогда тот, кто придёт по ссылке из статьи, узнает, что случилось.

Библиотека ведёт на мёртвый сайт. Классический набор рукописных цифр MNIST библиотека `torchvision` скачивала с личной страницы одного из авторов набора, Яна Лекуна. Страница начала отвечать отказом 403, и у пользователей ломалась загрузка. Разработчики библиотеки получили разрешение держать копию набора у себя, а позже поставили своё зеркало первым в списке. Спасло то, что у каждого файла в коде записана контрольная сумма: по ней видно, что файл с зеркала тот же самый. Croissant 1.1 настойчиво рекомендует записывать такие суммы для всех файлов набора, у которого есть версии.

Тестовая часть утекла в обучение. Языковые модели обучают на текстах из интернета, и опубликованные тестовые наборы попадают в обучение вместе с ответами. Тогда результат модели на бенчмарке завышен, и сравнение теряет смысл. Защищаются двумя способами. Первый — не публиковать ответы тестовой части: в наборе GLUE на Hugging Face у всех 1063 строк тестовой части задачи CoLA метка `-1`, а оценку считает сервер бенчмарка. Так же устроены скрытые таблицы результатов на Kaggle и Codabench. Второй — метка-«канарейка»: в каждый файл заданий бенчмарка BIG-bench вписана особая строка, чтобы составители обучающих наборов могли такие файлы найти и выбросить.

Файл только после входа. Kaggle в разметке страницы набора сам помечает файл как требующий входа (`requiresSubscription: true`). У Hugging Face автор может включить запрос доступа, у AI Hub данные получают только граждане Кореи, у AIKosh — после регистрации. Для человека это неудобство, для поисковика данных и сборщика — набор без файлов. Если ограничение нужно, карточка всё равно должна быть открыта, а причина и порядок доступа — написаны в ней.

Площадка закрылась. Papers with Code годами связывал статьи, код, наборы и таблицы результатов. 24 июля 2025 года Meta закрыла сервис, и адрес раздела наборов теперь ведёт на страницу популярных статей Hugging Face, где нет ни наборов, ни таблиц результатов. Последний снимок данных сохранили как архив, но ссылки из статей на страницы наборов перестали работать. Реестр каталогов узнаёт о таком не сразу: в выпуске 1.27.0 запись Papers with Code всё ещё помечена действующей и с API. Так же при закрытии CodaLab таблицы результатов не перенесли на Codabench. Вывод для автора набора: карточка на хабе — способ найти набор, а не место его долгого хранения. Копию выпуска с DOI держат в научном репозитории.

Страница пустая без скрипта. Раздел наборов на портале ИИ Москвы 12 октября 2026 года отдавал роботу пустую оболочку страницы около килобайта, а список наборов рисовал скрипт. Поисковик, веб-архив и программа-помощник такого списка не увидят. Как это лечится — в главе о находимости.

Свой сайт без API. У 78 самописных ML-каталогов реестра программный интерфейс отмечен у 13. Для ML-каталога это особенно странно: его пользователи работают из кода и браузером почти не пользуются.

Как проверить ML-каталог и набор

Шесть вопросов к конкретному набору:

1. **Есть ли машиночитаемое описание?** Запросите Croissant или найдите разметку schema.org на странице набора.
2. **Как набор разбит на части и где ответы тестовой части?** Если ответы опубликованы, результаты на нём могут быть завышены.
3. **Какая лицензия и откуда данные?** Лицензия должна быть названа, а для производного набора — указан исходный набор и его лицензия.
4. **Можно ли скачать без входа и без браузера?** Попробуйте загрузить набор из кода или по прямой ссылке.

5. **Есть ли версия и постоянный идентификатор?** DOI или хотя бы номер версии и контрольные суммы файлов.
6. **Есть ли копия вне площадки?** Если набор лежит только на одном хабе, его закрытие равно удалению.

```
# Hugging Face: описание Croissant и список файлов Parquet по частям набора
curl -s "https://huggingface.co/api/datasets/nyu-mll/glue/croissant"
curl -s "https://datasets-server.huggingface.co/parquet?dataset=nyu-mll/glue"
# Прочитать часть набора прямо с хаба, без скачивания архива (DuckDB 0.10.3 и новее):
# у тестовой части CoLA все метки равны -1, ответы скрыты
duckdb -c "SELECT label, count(*) FROM 'hf://datasets/nyu-mll/glue/cola/test-*.parquet' GROUP
BY 1"
# Kaggle: Croissant набора (ответ — переадресация на файл в облачном хранилище)
curl -sL "https://www.kaggle.com/datasets/unsdsn/world-happiness/croissant/download"
# OpenML: Croissant и описание набора с адресами файлов ARFF и Parquet
curl -s "https://www.openml.org/croissant/dataset/61"
curl -s "https://www.openml.org/api/v1/json/data/61"
# UCI: описание набора, в ответе есть DOI
curl -s "https://archive.ics.uci.edu/api/dataset?id=53"
# Dataverse: полное описание в Croissant
curl -s "https://dataverse.harvard.edu/api/datasets/export?
exporter=croissant&persistentId=doi:10.7910/DVN/TJCLKP"
# DataCite: сколько DOI наборов выдал Kaggle (префикс 10.34740); у Hugging Face префикс
10.57967
curl -s "https://api.datacite.org/dois?prefix=10.34740&resource-type-
id=dataset&page%5Bsize%5D=0"
# Проверить свой файл Croissant перед публикацией
pip install mlcroissant
mlcroissant validate --jsonld metadata.json
```

Все команды проверены 12 октября 2026 года. В ответах Hugging Face, Kaggle, OpenML и Dataverse поле `conformsTo` указывает Croissant 1.0. Общий список проверок — в чек-листах.

Как выбрать

Ситуация	Что выбрать	Почему
Набор к статье или бенчмарк исследовательской группы	Hugging Face или Kaggle, копия выпуска с DOI в Zenodo или научном репозитории	там набор ищут и загружают из кода; конференции принимают эти площадки, а копия с DOI переживёт площадку
Соревнование со скрытой тестовой частью	Codabench (общий codabench.org или своя установка) или соревнования Kaggle	скрытые ответы и таблица результатов есть в ядре; свой сайт этого не даст
Данные нельзя отдать на чужую площадку: медицина, персональные данные	свой репозиторий с доступом по заявке, например Dataverse (Croissant в ядре с версии 6.10), или защищённая среда	закрытые файлы с запросом доступа есть в ядре Dataverse; открытая карточка в Croissant делает набор находимым
Портал открытых данных хочет, чтобы его наборы брали для обучения	не строить отдельный ML-каталог; включить плагин `croissant` в CKAN, отдавать таблицы в Parquet со словарём полей	разработчики найдут набор через поисковик данных и загрузят без ручной работы
Коллекция лаборатории из десятка наборов	статический сайт с файлом Croissant у каждого набора, файлы на хабе или в репозитории с DOI	дёшево и долговечно; так сделаны большинство сайтов бенчмарков
Государственная программа данных для ИИ	каталог с открытыми карточками и доступом по заявке, Croissant, выгрузка описаний и API	закрытые данные остаются под контролем, а описания видны снаружи и попадают в поисковики

Когда свой ML-каталог не нужен

Почти всегда, если цель — чтобы набор нашли и использовали. Аудитория, загрузка из кода и Croissant уже есть на хабах, а свой сайт придётся годами сопровождать ради нескольких наборов. Достаточно аккуратной карточки на хабе, названной лицензии, разбиения на части и копии выпуска с DOI.

Свой каталог оправдан в трёх случаях: данные нельзя выкладывать на чужую площадку; хаб не принимает ваш вид файлов или объём; вы — государственная или отраслевая программа, которая должна учитывать и раздавать наборы многих организаций. Даже тогда карточку стоит продублировать там, где набор будут искать. Как обосновать такое решение, разобрано в главе «Своя разработка или готовое решение».

Чем отличается от соседей

Сосед	Главное отличие
Научный репозиторий	главное — неизменная версия с DOI и ссылка из статьи; в ML-каталоге — загрузка в код и сравнение моделей. Копию выпуска ML-набора часто кладут в Zenodo именно ради DOI и сохранности
Портал открытых данных	набор не разбит на части, у описания нет раздела о перекосах; публикует орган власти, а не автор набора
Каталог микроданных	данные о людях из обследований с доступом по заявке. Государственные платформы данных для ИИ со «безопасными зонами» устроены похоже
Маркетплейс	наборы продают; продавцы обучающих данных вроде Datatang — скорее маркетплейсы, чем ML-каталоги
Git-репозиторий	многие бенчмарки живут в репозитории на GitHub с сайтом на GitHub Pages; это публикация без портала, о ней — глава о Git-репозиториях
Реестр моделей внутри компании	учёт своих моделей для разработки и выпуска, например в MLflow; это корпоративный инструмент, а не публичный каталог

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: число ML-каталогов, владельцы, страны, платформы, доля API и статусы. Разметка записей по видам сделана для книги вручную по списку записей. Запросы — в приложении.
- Число наборов и моделей: Hugging Face Datasets и модели, API OpenML, API архива UCI — 12 октября 2026 года
- Croissant: сайт и список площадок с поддержкой, спецификация 1.1 от 29 января 2026 года, расширение RAI 1.0, GeoCroissant 1.0, библиотека mlcroissant и валидатор
- Карточки: T. Gebru и др., Datasheets for Datasets, 2018; M. Mitchell и др., Model Cards for Model Reporting, 2019; карточки наборов на Hugging Face
- NeurIPS 2026: правила трека Evaluations & Datasets, размещение наборов и требования к полям RAI, 4 мая 2026 года
- Hugging Face: доступ по запросу, DOI, лимиты хранения, перевод в Parquet и его предел 5 ГБ, Croissant для наборов, DuckDB и адреса hf://

- Kaggle: документация по наборам — лимит 200 ГБ на набор; Croissant и разметка страницы набора — проверка 12 октября 2026 года
- Свои установки: Croissant в ckanext-dcat и выпуски расширения; Croissant в Dataverse и прежний внешний экспортёр; форматы выгрузки InvenioRDM в исходном коде — Croissant среди них нет; OpenML, Codabench
- CodaLab и Codabench: объявление о закрытии CodaLab, переход на Codabench, итоги 2025 года
- DOI ML-площадок: API DataCite — префиксы 10.34740 (Kaggle), 10.57967 (Hugging Face), 10.24432 (UCI), 12 октября 2026 года
- Закон ЕС об ИИ, регламент 2024/1689; шаблон сводки обучающих данных, 24 июля 2025 года, и ответы на вопросы о нём
- Государственные платформы: условия использования AI Hub и безопасная зона; AIKosh; NAIRR Pilot; AI-on-Demand; реестр ЦИТ и сообщение о его обновлении, декабрь 2025 года; число наборов, лицензии в карточках реестра ЦИТ и HTML раздела наборов ai.mos.ru — проверка 12 октября 2026 года
- S. Longpre и др., A large-scale audit of dataset licensing and attribution in AI, Nature Machine Intelligence, 2024; препринт 2023 года
- X. Yang и др., Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face, ICLR 2024
- LAION: снятие LAION-5B, 19 декабря 2023 года; выпуск Re-LAION-5B, 30 августа 2024 года
- MNIST в torchvision: загрузчик и зеркала, переход на своё зеркало, ошибка 403 на исходном сайте
- BIG-bench — метка-«канарейка» в файлах заданий; набор GLUE на Hugging Face — проверка тестовой части CoLA 12 октября 2026 года
- Закрытие Papers with Code — обсуждение перенаправления на Hugging Face; адрес paperswithcode.com/datasets ведёт на huggingface.co/papers/trending — проверка 12 октября 2026 года
- И. Бегтин, Telegram: «сначала код» в ML-каталогах, 21.03.2024; AIKosh, 26.08.2026; национальные каталоги для ИИ, 30.09.2026
- Облачные хранилища и общие хабы, ИИ, пространства данных и регулирование

Каталоги метаданных и реестры

Каталог метаданных описывает не сами данные, а то, чем данные описывают и где их искать: словари терминов, справочники кодов, определения полей, структуры статистических таблиц, описания наборов, которые лежат в другом месте, и перечни других каталогов. На такой сайт приходят не скачать таблицу, а узнать, какой код у темы «Транспорт», что значит поле «специальность автора» или какие репозитории принимают данные о климате.

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) таких каталогов 172 — 0,4% всех каталогов. Тип маленький, но все записи действующие, а программный доступ есть у 67% — намного больше, чем у каталогов показателей (37%), и почти столько же, сколько у порталов открытых данных (69%).

Главный вывод главы: ценность каталога метаданных — в том, что на него ссылаются другие. Поэтому у каждой записи должен быть постоянный адрес, версия и отметка, чем её заменили. Описания должны отдаваться программе в стандарте (SKOS, SDMX, DCAT), а вся база — скачиваться целиком. Реестр, который читают только глазами, через несколько лет перестают цитировать, и он превращается в ещё один сайт со списком.

Как устроена запись

В портале открытых данных главное — файл, в каталоге показателей — число с разрезами. В каталоге метаданных единица — описание, на которое ссылаются: термин, код, элемент данных, описание набора или каталога. Три живых примера показывают, как это работает.

Термин как адрес. В словарях Публикационного офиса ЕС у темы «Транспорт» есть адрес `http://publications.europa.eu/resource/authority/data-theme/TRAN`. По нему программа получает описание темы в RDF с названиями на языках ЕС (проверено 12 октября 2026 года). Порталы по профилю DCAT-AP пишут в поле темы не слово «Транспорт», а этот адрес. Поэтому data.europa.eu складывает транспортные наборы разных стран, не переводя и не угадывая слова. Так работает любой хороший словарь: термин — это адрес, а название — одна из его подписей.

Код из общего справочника. В ключе ряда курса доллара Европейского центрального банка `EXR.D.USD.EUR.SP00.A` буква `D` — не выдумка банка. Это код «ежедневно» из справочника частот `CL_FREQ`, который лежит в глобальном реестре SDMX. На 12 октября 2026 года в этом справочнике версии 2.1 34 кода: `A` — год, `Q` — квартал, `M` — месяц, `W` — неделя, `D` — день и другие. Справочник один на все статистические службы, поэтому `M` значит «ежемесячно» и у Евростата, и у МВФ. Как устроен ключ ряда, разобрано в главе о каталогах показателей.

Элемент данных с паспортом. Швейцарская платформа I14Y описывает понятия, справочники, наборы и сервисы органов власти. Первое понятие в её открытом API 12 октября 2026 года — поле `DocumentEntry.author.authorSpeciality` от eHealth Suisse: специальность автора медицинского документа. У него есть описание на четырёх языках, тип (справочник), тип значения (строка), предельная длина (30 знаков), издатель, версия, статус регистрации, дата начала действия и поля «заменяет» и «заменён». Именно такой паспорт поля описывает стандарт ISO/IEC 11179, о нём — ниже.

Что получает свою страницу

В зависимости от вида каталога свой адрес получают:

1. **Понятие или код** — одна тема, одна рубрика, один код справочника.
2. **Словарь или справочник целиком** — тезаурус, классификатор, онтология, кодовый список.
3. **Элемент данных** — определение поля с допустимыми значениями.
4. **Структура** — описание таблицы: какие у неё разрезы и какие справочники к ним привязаны.
5. **Описание набора без данных** — что есть у организации и как это получить.
6. **Описание каталога, стандарта или инфраструктуры** — карточка репозитория, словаря, сети станций наблюдения.

Число записей в каталоге метаданных поэтому нельзя сравнивать с числом наборов на портале. В глобальном реестре SDMX 12 октября 2026 года было 451 кодовый список, 248 структур таблиц и 1167 описаний потоков данных вместе со всеми версиями — и при этом ни одного значения статистики.

Кто создаёт

Каталоги метаданных почти поровну делят наука, международные организации и центральные органы власти.

Владелец	Каталогов	Доля
Университеты и научные организации	54	31%
Международные организации	52	30%
Центральные органы власти	41	24%
Регионы	8	5%
Общественные организации	6	3%
Компании	6	3%
Сообщества и прочие	5	3%

- **Научные организации и библиотеки** ведут словари и онтологии: Национальная библиотека Финляндии — сервис словарей Finto, EMBL-EBI — поиск по онтологиям наук о жизни OLS, Стэнфорд — BioPortal. Медицинские центры и сети исследований редких болезней ставят точки FAIR Data Point, чтобы рассказать о своих когортах и регистрах пациентов.
- **Международные организации** держат общие справочники и структуры: ФАО — тезаурус AGROVOC, ЮНЕСКО — свой тезаурус, Банк международных расчётов — глобальный реестр SDMX, Всемирная метеорологическая организация — реестры станций и спутников наблюдения OSCAR, секретариат Конвенции о биологическом разнообразии — механизмы обмена информацией.
- **Центральные органы власти** ведут реестры элементов данных и справочников: Австралийский институт здоровья и благосостояния — METEOR, правительство Нового Южного Уэльса — реестр метаданных, Швейцария — I14Y, нидерландская служба Logius —

Stelselcatalogus, каталог понятий базовых государственных реестров, Библиотека Конгресса США — службу связанных данных id.loc.gov.

По странам первые места — Нидерланды (24), Великобритания (20), Испания (9), США и Германия (по 7), Австралия (6); ещё 40 каталогов охватывают весь мир и 11 — Евросоюз. Как и в других типах, число говорит о том, какое ПО прижилось у одного сообщества, а не о зрелости страны. В Нидерландах 19 из 24 — точки FAIR Data Point медицинских центров и исследовательских сетей. В Великобритании 15 из 20 — каталоги на облачном сервисе Metadata Browser. В Испании 7 из 9 — тоже точки FAIR Data Point, шесть из них у банков семян ботанических садов и университетов. В Австралии 4 из 6 — реестры на Aristotle Metadata Registry.

Граница типа размыта, и 172 — нижняя оценка. В реестре нет самого re3data, Research Vocabularies Australia записан научным репозиторием, а ирландский каталог охраняемых данных — порталом открытых данных на CKAN. Зато сюда попали каталоги других объектов: программных пакетов (Libraries.io, Google Open Source Insights), журналов (портал ISSN), товарных кодов (WTO HS Tracker). Они тоже описывают, а не хранят, но к публичным данным относятся косвенно.

Что публикуют

- **Словари, тезаурусы, классификаторы и онтологии.** Темы, предметные рубрики, коды стран и языков, виды деятельности, термины науки. По ним другие каталоги заполняют поля одинаково. Онтология отличается от словаря строгостью: кроме связей «шире — уже» она задаёт классы, свойства и правила, по которым программа делает выводы.
- **Структуры и элементы данных.** Какие разрезы есть у показателя и какие коды допустимы в каждом. Что значит поле «дата регистрации», в каком формате его пишут, кто отвечает за определение и какая редакция сейчас действует.
- **Схемы.** Шаблоны описания: какие поля должны быть у записи о наборе, образце или объекте. Так работают реестры схем CEDAR и buildingSMART Data Dictionary для строительства.
- **Описания наборов без самих данных.** Организация рассказывает, что у неё есть, а сами данные остаются у неё: медицинские когорты, регистры пациентов, коллекции биобанков, охраняемые данные государственного сектора.
- **Перечни каталогов, стандартов и инфраструктур.** Карточка репозитория, словаря или стандарта: адрес, тематика, условия, сертификаты, интерфейсы.

Чего в каталоге метаданных обычно нет: самих значений, файлов и записей. Если сайт даёт скачать таблицу, это уже портал или каталог показателей, даже если таблица — справочник. Об этой границе — в разделе о порталах справочников.

Стандарты

SKOS (W3C, 2009) — способ описать словарь в RDF. У каждого понятия есть адрес, основное название и синонимы на разных языках, код, связи «шире — уже» и «связано с», а также соответствия понятиям других словарей. Почти все открытые словари отдают в SKOS. Для

статистических классификаций есть расширение **XKOS** (DDI Alliance, версия 1.0 от 2019 года): уровни классификации, таблицы соответствия между редакциями, отличия «включает» и «не включает».

OWL — язык онтологий W3C. В нём описывают области знания с классами и правилами: болезни, гены, виды растений, сельскохозяйственные культуры. Каталоги онтологий — OLS, BioPortal, AgroPortal — принимают онтологии в OWL и показывают их людям и программам.

ISO/IEC 11179 «Реестры метаданных» — как описать отдельный элемент данных. Элемент состоит из понятия (объект и свойство: «человек — пол») и представления (допустимые значения: код из списка, дата, число с единицей). У элемента есть определение, хранитель, отвечающий за него, и статус регистрации — от кандидата до стандарта, а потом «заменён» или «выведен». Действующая редакция метамодели — ISO/IEC 11179-3:2023, в 2026 году к ней вышла поправка о сопоставлении элементов. Стандарт используют там, где разные системы должны одинаково понимать поле: в здравоохранении, государственной статистике, обмене между ведомствами.

SDMX — модель структур статистики: кодовые списки, схемы понятий, структуры таблиц, потоки данных. Реестр SDMX хранит эти структуры, чтобы все участники обмена брали их из одного места. Подробно стандарт разобран в главе о каталогах показателей.

DCAT и его профили. FAIR Data Point отдаёт описания каталогов и наборов в DCAT, поэтому их забирают те же сборщики, что и у порталов открытых данных. Для медицинских данных Еврокомиссия разрабатывает профиль **HealthDCAT-AP**: версия Release 7 опубликована в мае 2026 года как проект спецификации. Он добавляет к DCAT-AP поля, которые нужны для данных о здоровье: орган доступа к данным, категория медицинских данных, число записей и число уникальных людей, возрастной охват, срок хранения, цель обработки и правовое основание.

Описание репозиториев. У re3data своя схема (версия 4.0): тематика, типы данных, условия доступа к данным и к самому хранилищу, сертификаты, интерфейсы. По ней же в главе о классификации описаны вопросы о доступе.

Для страниц, которые читают поисковики, у schema.org есть типы `DefinedTermSet` и `DefinedTerm` — словарь и термин. Это не замена SKOS, а способ сделать словарь находимым.

Каталоги описаний по закону

В Евросоюзе каталоги описаний закрытых данных из добровольного дела становятся обязательным.

- **Акт об управлении данными** (Data Governance Act, регламент 2022/868, статья 8) обязывает каждую страну ЕС завести единую информационную точку. В ней должен быть доступный для поиска перечень охраняемых данных государственного сектора — персональных, коммерчески тайных, статистически конфиденциальных — с форматом, объёмом и условиями повторного использования. Через точку подают запрос на доступ. Еврокомиссия собирает национальные перечни в Европейский реестр охраняемых данных государственного сектора (ERPD) на data.europa.eu.

- **Регламент о Европейском пространстве медицинских данных** (EHDS, регламент 2025/327, статья 77) требует от органов доступа к медицинским данным вести публичный машиночитаемый каталог описаний наборов, доступных для вторичного использования: источник, охват, основные свойства и условия доступа. Каталог должен быть доступен единой информационной точке по Акту об управлении данными. Правила вторичного использования для большинства категорий данных действуют с 26 марта 2029 года, для пяти категорий, включая геномные данные, — с 26 марта 2031 года. Проект исполнительного акта Еврокомиссии 2026 года предлагает описывать наборы по HealthDCAT-AP.

Для разработчика это значит: каталог описаний закрытых данных должен отдавать DCAT-AP или его отраслевой профиль и быть доступен для сбора, иначе он не попадёт в национальную точку и европейский реестр.

Платформы

Своя разработка — 38% (65 каталогов), меньше, чем у каталогов показателей (44%). Программный доступ у самописных каталогов метаданных есть только у 17%, а у специализированных платформ — почти у всех.

Платформа	Каталогов в реестре	Что это	Программный доступ
Своя разработка	65	реестры каталогов, словари, справочники ведомств	у 17%
FAIR Data Point	47	свободная (MIT) точка описаний организации в DCAT; выпуск 1.22.2 — 24 сентября 2026 года	у всех: DCAT в Turtle, описание API по OpenAPI
Metadata Browser	16	коммерческий облачный сервис британской компании MetadataWorks для каталогов медицинских и административных данных	у всех
Skosmos	10	свободная (MIT) программа публикации словарей SKOS Национальной библиотеки Финляндии; выпуск 3.4 — 31 августа 2026 года	REST API в ядре
Aristotle Metadata Registry	8	реестр элементов данных по ISO/IEC 11179; облачный сервис	API у всех
Fusion Metadata Registry	7	реестр структур SDMX Банка международных расчётов; бесплатное приложение, выпуск 12.4.2 — 19 сентября 2026 года	SDMX REST API в ядре
MOLGENIS	5	свободная (LGPL-3.0) платформа из Гронингена; каталоги биобанков и когорт	GraphQL у всех
OntoPortal	2	свободная (BSD) платформа каталогов онтологий; на ней BioPortal и AgroPortal	REST API
Отраслевые системы	7	WMO OSCAR, механизмы обмена Конвенции о биоразнообразии, справочники SeaDataNet	у большинства
Другие	5	по одному: DataHub, EntryScape, PublishMyData, CEDAR, Drupal	—

Видно, что тип делится на три мира. Первый — узкие свободные инструменты, каждый под один вид описаний: Skosmos и OntoPortal — для словарей и онтологий, Fusion Metadata Registry — для структур SDMX, FAIR Data Point — для описаний наборов в DCAT. У них API и

стандартный формат в ядре. Второй — облачные реестры: Aristotle для элементов данных и Metadata Browser для каталогов чувствительных данных. Удобно, но описания живут у поставщика. Третий — собственные разработки ведомств и проектов, где машинный доступ — скорее исключение.

Некоторые известные установки реестр записывает как свою разработку, хотя за ними стоит тиражируемое ПО. Например, AGROVOC отвечает через стандартный REST API Skosmos, а поиск по онтологиям OLS — это свободная программа EMBL-EBI (Apache 2.0, выпуск 4.0.3 — 8 октября 2026 года), которую можно поставить у себя.

Судьба открытого кода Aristotle

Aristotle Metadata Registry начинался как свободная программа на Django, построенная по ISO/IEC 11179. В августе 2020 года из открытого репозитория убрали код: там остался только файл с объявлением, что продукт перерабатывают и предлагают как облачный сервис. Те, кто поставил открытую версию у себя, остались с кодом без обновлений. Все восемь реестров на Aristotle в выгрузке реестра — облачные. Перед выбором платформы для реестра проверяют, кто её ведёт и когда выходил последний выпуск, — так же, как у расширений.

Ядро и дополнения

У платформ каталогов метаданных нет экосистемы расширений, как у CKAN. Каждая делает одно дело, а недостающее добавляют соседним продуктом.

Возможность	FAIR Data Point	Skosmos	Fusion Metadata Registry	Aristotle (облако)
Что хранит	описания каталогов, наборов и дистрибутивов	словари SKOS	структуры SDMX: справочники, понятия, таблицы	элементы данных по ISO/IEC 11179
Машинный формат	DCAT в RDF (Turtle и другие)	SKOS через REST API	SDMX-ML, SDMX-JSON	свой API
Правка в интерфейсе	ядро	нет: словарь готовят в другом редакторе и загружают в хранилище SPARQL	ядро	ядро
Версии и статусы	ядро	нет, версии ведут в источнике словаря	ядро: у каждой структуры версия	ядро: статусы регистрации
Обработка данных	нет	нет	проверка и перевод данных между форматами SDMX	нет
Регистрация в общем каталоге	ядро: точка сообщает о себе в индекс FAIR Data Point	нет	нет	нет
Как разворачивают	контейнер на своём сервере	веб-приложение на своём сервере поверх хранилища SPARQL	контейнер на своём сервере	только у поставщика

Если нужен каталог каталогов, а не словарь или реестр структур, специальное ПО не обязательно. Каталог экосистемы CKAN сам работает на CKAN, а реестр dataportals-registry — это набор файлов YAML в Git-репозитории с выгрузками. Для описаний закрытых данных

государственного сектора тоже часто берут обычный портал: ирландский каталог охраняемых данных сделан на CKAN.

Как выбрать

Выбор начинается с вопроса, что именно вы описываете и кто будет ссылаться на ваши записи.

Ситуация	Что взять	Почему и чем платим
Нужно опубликовать свой словарь или классификатор	SKOS с постоянными адресами понятий; витрина — Skosmos; маленький словарь — файл SKOS в Git и постоянные адреса через w3id.org	стандартный формат читают все; Skosmos только показывает словарь, редактор нужен отдельно
Отрасль ведёт много онтологий	свой экземпляр OntoPortal, как AgroPortal, или регистрация онтологий в BioPortal и OLS	готовый поиск и сопоставления между онтологиями; свой экземпляр надо сопровождать
Статистической службе нужны общие справочники и структуры	Fusion Metadata Registry или реестр в составе .Stat Suite	SDMX в ядре, проверка данных по структурам; внедрение требует знания SDMX
Ведомства должны одинаково понимать поля	реестр элементов данных по ISO/IEC 11179: облачный Aristotle или национальная платформа вроде I14Y	паспорт поля со статусами и версиями; облачный сервис держит описания у поставщика
Организация хочет рассказать о закрытых данных	FAIR Data Point или портал с выгрузкой DCAT-AP; для медицины в ЕС — с полями HealthDCAT-AP	описания собирают национальная точка и европейский реестр; сами данные остаются у владельца
Нужен перечень каталогов по теме или стране	сначала проверить re3data, FAIRsharing, BARTOC и dataportals-registry; если своего перечня не избежать — файлы в Git с выгрузкой CSV и JSON	перечень, который можно скачать, проверят и переиспользуют другие

Своя разработка оправдана, когда описания сложнее, чем умеют готовые продукты, и есть команда, которая будет вести их годами. Образцы — I14Y и id.loc.gov: у обоих открытый API и постоянные адреса записей. Если команда не готова сделать машинный доступ, свой реестр повторит судьбу самописных каталогов метаданных, из которых программы читают только 17%. Общая логика выбора — в главе «Свои силы, свободное ПО или облачный сервис».

Когда свой каталог не нужен

Большинству организаций свой каталог метаданных не нужен. Словарь можно опубликовать файлом SKOS в Git-репозитории и зарегистрировать в BARTOC или Linked Open Vocabularies. Репозиторий — внести в re3data, стандарт или базу данных — в FAIRsharing, каталог любого типа — в dataportals-registry. Описания своих наборов достаточно отдавать в DCAT с портала, на котором они уже лежат. Отдельный реестр нужен тому, на кого будут ссылаться многие: статистической службе, национальной библиотеке, отраслевой инфраструктуре, органу, которому закон велит вести перечень. Способы публикации без портала — в части «Публикация без портала».

Что отличает полезный каталог метаданных

Шесть признаков, по которым видно, что на каталог можно опереться.

1. **Постоянные адреса.** У каждого понятия, кода и описания свой адрес, который не меняется при переезде сайта. Адрес темы `data-theme/TRAN` в словарях ЕС одинаков для всех порталов Европы.
2. **Версии и замены.** Устаревшая запись не удаляется, а помечается, и видно, чем её заменили. В I14Y у каждого понятия есть версия, статус, дата начала действия и ссылки «заменяет» и «заменён».
3. **Машинный доступ в стандарте.** Запись отдаётся программе в SKOS, SDMX или DCAT, а не только страницей. Хороший признак — один и тот же адрес отдаёт человеку страницу, а программе RDF или JSON в зависимости от заголовка запроса.
4. **Выгрузка целиком.** Весь словарь или весь перечень скачивается одним файлом. re3data отдаёт список всех репозиториев одним запросом, реестр dataportals-registry — одним файлом.
5. **Понятный хранитель.** У каждой записи указан тот, кто отвечает за определение и правки, а у каталога — правила включения записей. В нидерландском Stelselcatalogus за описания отвечают держатели самих реестров, и они договорились о сроках и полноте поставки.
6. **Регистрация уровнем выше.** Словарь записан в BARTOC, онтология — в OLS или BioPortal, точка FAIR Data Point — в индексе точек, каталог закрытых данных — в национальной единой точке.

Примеры

Словари и онтологии

- Словари Публикационного офиса ЕС — темы, типы файлов, языки, страны, права доступа и другие справочники, на которые опирается DCAT-AP. На них же ссылаются наборы и сервисы швейцарской I14Y.
- id.loc.gov — служба связанных данных Библиотеки Конгресса США: предметные рубрики, имена, языки. Любая рубрика отдаётся в JSON-LD, если добавить к её адресу `.json`.
- AGROVOC — многоязычный тезаурус ФАО по сельскому хозяйству и питанию, опубликован на Skosmos.
- Finto — национальный сервис словарей Финляндии на Skosmos, от его команды и пошла сама программа.
- Ontology Lookup Service — поиск по онтологиям наук о жизни EMBL-EBI: 287 онтологий на 12 октября 2026 года.
- BioPortal и AgroPortal — каталоги онтологий медицины и сельского хозяйства на OntoPortal.
- Research Vocabularies Australia — общий каталог научных словарей Австралии.

Структуры и элементы данных

- Глобальный реестр SDMX — общие структуры и справочники международных организаций на Fusion Metadata Registry. На той же программе работают реестры Евростата, МВФ, ЮНИСЕФ, статистического центра стран Персидского залива и статистической службы Таиланда.

- METEOR — реестр элементов данных здравоохранения и социальной сферы Австралии, который ведёт Австралийский институт здоровья и благосостояния.
- I14Y — швейцарская платформа совместимости: понятия, справочники, наборы, сервисы данных и государственные услуги с открытым API.
- Stelselcatalogus — каталог понятий, элементов данных и информационных моделей базовых государственных реестров Нидерландов: населения, торгового реестра, кадастра и других. Обновлён в июне 2026 года; описания наборов и сервисов при этом остались на data.overheid.nl и developer.overheid.nl.
- schema.gov.it — национальный каталог онтологий, словарей и схем данных Италии, подробнее — ниже.

Описания наборов без данных

- Точки FAIR Data Point нидерландской сети Health-RI, медицинских центров Лейдена и Неймегена, европейской программы по редким болезням.
- Каталог BBMRI-ERIC — европейский перечень биобанков и их коллекций на MOLGENIS.
- Каталоги на Metadata Browser: Управления национальной статистики Великобритании, программы ADR UK для исследований административных данных, Genomics England, региональных доверенных исследовательских сред NHS.
- Ирландия и UK Data Service — в разделе об открытых описаниях закрытых данных.

Каталоги каталогов

Каталог каталогов отвечает на вопрос «где вообще искать данные по этой теме или стране». Он описывает порталы, репозитории или словари, а не наборы. Поисковиком он не является: искать придётся уже на найденном портале. Зато он помогает, когда общий поисковик ничего не дал, потому что нужный каталог в него просто не попал.

Каталог	Кто ведёт	Что описывает	Сколько записей
dataportals.org, прежде datacatalogs.org	Open Knowledge Foundation	порталы открытых данных; список выгружается в CSV и JSON	631, октябрь 2026
Open Data Inception	Huwise (бывшая OpenDataSoft)	порталы открытых данных по странам	на сайте «2600+», в исходном наборе 3763 записи, октябрь 2026
datacatalogs.ru	Инфокультура	каталоги и репозитории данных России и о России: тип, темы, ПО, наличие API, режим доступа	245 при запуске в ноябре 2021, 263 в июне 2022
re3data	DataCite и партнёры	научные репозитории	3535, 12 октября 2026
FAIRsharing	группа Оксфордского университета	стандарты, базы данных и политики журналов и фондов	1973 базы, 1646 стандартов, 159 политик, январь 2023
BARTOC	библиотечный центр VZG, Германия	словари, тезаурусы, классификации и онтологии	4162, октябрь 2026
Linked Open Vocabularies	Мадридский политехнический университет	словари связанных данных: DCAT, schema.org, FOAF и другие	920, октябрь 2026
Индекс FAIR Data Point	сообщество FAIR Data Point	точки FAIR Data Point, которые сами о себе сообщили	230, из них активны 74, октябрь 2026

Каталог	Кто ведёт	Что описывает	Сколько записей
ecosystem.ckan.org	сообщество CKAN	расширения и порталы на CKAN с версиями ядра	около 1,4 тысячи расширений и 1,3 тысячи порталов, октябрь 2026
dataportals-registry	проект Dateno	каталоги всех типов, по YAML-файлу на каталог	больше 43 тысяч, выпуск 1.27.0, октябрь 2026

Числа в таблице несравнимы между собой: каждый перечень считает своё и по-своему решает, что считать каталогом. Почему число каталогов в мире зависит от усилий поиска, разобрано в главе «Мир каталогов в цифрах». О каталоге экосистемы CKAN подробнее — в главе о CKAN.

Открытые описания закрытых данных

Описание набора может быть открытым, даже когда сами данные закрыты. Такой каталог честно говорит: данные есть, вот что в них, вот как получить доступ, если у вас есть основание. Без него человек не может запросить данные просто потому, что не знает об их существовании. В ЕС такие каталоги теперь требует закон — см. раздел «Каталоги описаний по закону».

- Каталог данных государственного сектора Ирландии описывает данные органов власти, которые нельзя опубликовать открыто: персональные, коммерческая тайна, статистическая конфиденциальность. По Акту об управлении данными это «охраняемые данные»: их можно получить для повторного использования на условиях, которые сохраняют защиту. Каталог служит единой точкой, где ведомства перечисляют такие данные, а желающий подаёт запрос. На 12 октября 2026 года в нём 1543 набора.
- UK Data Service — британский архив социальных и экономических данных — держит три уровня доступа. Открытые данные скачиваются без регистрации по открытым лицензиям. Данные под охраной (safeguarded) — после регистрации и согласия с пользовательским соглашением. Контролируемые данные — подробные и конфиденциальные — доступны только опытным исследователям по заявке и только внутри доверенной исследовательской среды SecureLab: с одобренного рабочего места, без права скачать сами данные. При этом в 2023 году в общем открытом каталоге были перечислены наборы всех трёх уровней, а документация к закрытым была открытой.
- Точки FAIR Data Point устроены так же, но для науки: медицинский центр перечисляет когорты и регистры пациентов в DCAT, а доступ к самим записям обычно получают по заявке.

Подробнее о каталогах микроданных — в отдельной главе, о лицензиях и режимах доступа — в главе о лицензиях и доступе.

Порталы справочников: пограничный случай

Нормативно-справочная информация (НСИ) — это справочники и классификаторы, которые используют информационные системы: коды, перечни, словари значений. В России для неё есть отдельные государственные порталы: единая система НСИ на Госуслугах, портал НСИ

Минздрава, справочники ФОМС, системы НСИ Москвы и Санкт-Петербурга. В июне 2024 года на портале Минздрава было около 1,8 тысячи справочников с выгрузкой в XML, JSON, XLS и CSV.

Внешне такие порталы похожи на порталы открытых данных: много наборов, машиночитаемая выгрузка, API. Отличие в том, что у них один владелец, единые условия использования вместо лицензии на каждый набор, и созданы они ради обмена между государственными системами, а не ради публикации. В других странах похожие ресурсы чаще оформлены как каталоги схем и словарей. Итальянский schema.gov.it, который ведут Департамент цифровой трансформации и статистическая служба ISTAT, — национальный каталог онтологий, словарей и схем данных в RDF с точкой запросов SPARQL. Швейцарская I14Y хранит справочники как понятия с версиями и статусами, а не как файлы.

Куда относить портал справочников, однозначного ответа нет. По содержимому это ближе к каталогу метаданных: справочник описывает значения, которыми заполняют другие данные. По устройству — к порталу открытых данных. Решающий вопрос — что человек делает на сайте: скачивает справочник как данные или сверяет по нему коды в своей системе.

Где ломается

Карточка жива, каталога нет. Реестр описывает каталог или словарь, которого уже нет: адрес сменился, сайт закрыт, а карточка жива. Для реестра каталогов это профессиональный риск. Индекс FAIR Data Point на 12 октября 2026 года знал 230 точек, но активными были только 74. Ещё 80 не отвечали, 47 значились неактивными, у 28 описание не прошло проверку. Точки часто заводят под проект, и после его окончания сервер выключают. Карточку нужно помечать как недействующую, а не удалять молча: на старый адрес могли сослаться.

Перечень смешивает каталоги с тем, что каталогом не является. В Open Data Inception в 2023 году среди записей встречались законы о данных, страницы политики и отдельные наборы. В октябре 2026 года на сайте всё ещё написано «2600+», хотя в исходном наборе 3763 записи, — счётчик и содержимое живут отдельно. Чем больше перечень, тем нужнее явный критерий включения и пометка о проверке.

Перечень нельзя скачать. datacatalogs.ru в 2022 году был сайтом поверх базы в Airtable: вести её вручную удобно, но сервис проприетарный, а выгрузки каталога открытыми данными не было. Перечень, который нельзя скачать целиком, другие проекты не могут проверить и переиспользовать. Реестр dataportals-registry, начатый в 2023 году, устроен наоборот: каждая запись — файл в открытом репозитории, весь перечень выгружается одним файлом.

Словарь меняется без версий. Код темы переименовали, и все каталоги, которые на него ссылались, теперь показывают пустое поле. У словаря должны быть постоянные адреса терминов и отметка, какой термин устарел и чем его заменили. Реестры по ISO/IEC 11179 и SDMX держат версии в самой модели, у простых словарей на сайте их часто нет вовсе.

Платформа уходит в облако. Открытый код Aristotle Metadata Registry убрали из репозитория в 2020 году, продукт развивается только как облачный сервис. Облачный реестр удобен, пока жив договор. Записать в договор выгрузку всех элементов в открытом формате

нужно до начала работы, а не при переезде.

Национальный перечень не доходит до общего. Европейский реестр охраняемых данных на data.europa.eu 12 октября 2026 года содержал 10 757 описаний, но только от 12 стран. 69% записей (7408) дала Норвегия, причём среди них обычные открытые геоданные — ортофотопланы и данные о рельефе с пометкой «открытый доступ». Ирландия, в собственном каталоге которой 1543 набора, передала туда 137 записей. Что считать охраняемыми данными и что передавать наверх, страны пока понимают по-разному, а сбор работает только у тех, кто отдаёт DCAT-AP по правилам сборщика.

Словарь закрыт для программ. Словарь нужен прежде всего программам, но сайты иногда отказывают им наравне с ботами. 12 октября 2026 года Finto и METEOR ответили на запросы программы проверки ошибкой 403, в том числе с заголовком обычного браузера, хотя REST API других установок Skosmos отвечал. Если каталог метаданных закрыт фильтром от роботов, им могут пользоваться только люди — а значит, на него не сошлётся ни один портал. Общий разбор — в антипаттерне «Дверь закрыта для программ».

Как проверить каталог метаданных

Четыре шага на четверть часа.

1. **Возьмите одну запись и откройте её адрес программой.** Отдаёт ли адрес RDF, JSON или SDMX, а не только страницу? Остаётся ли он прежним после переезда сайта, по веб-архиву?
2. **Найдите устаревшую запись.** Видно ли, что она устарела и чем её заменили, или её просто удалили?
3. **Скачайте весь каталог.** Есть ли выгрузка целиком или только постраничный поиск?
4. **Проверьте, где каталог зарегистрирован.** Есть ли он в BARTOC, re3data, FAIRsharing, индексе FAIR Data Point, и совпадает ли там адрес?

Типичные адреса у разных платформ (проверено 12 октября 2026 года):

```
# FAIR Data Point: описание точки в DCAT (Turtle)
curl -s -H "Accept: text/turtle" "https://fdp.healthdata.nl/"
# Индекс FAIR Data Point: все зарегистрированные точки и их состояние
curl -s "https://home.fairdatapoint.org/index/entries/all"
# Skosmos: список словарей установки
curl -s "https://data.ub.uio.no/skosmos/rest/v1/vocabularies?lang=en"
# OLS: сколько онтологий в каталоге
curl -s "https://www.ebi.ac.uk/ols4/api/ontologies?size=1"
# Реестр SDMX: справочник частот, из которого код D в ключе ряда ЕЦБ
curl -s "https://registry.sdmx.org/sdmx/v2/structure/codelist/SDMX/CL_FREQ/latest?
format=sdmx-json"
# Словари ЕС: тема «Транспорт» как адрес, ответ в RDF
curl -sL -H "Accept: application/rdf+xml"
"http://publications.europa.eu/resource/authority/data-theme/TRAN"
# I14Y: понятия с версиями и статусами регистрации
curl -s "https://api.i14y.admin.ch/api/public/v1/concepts?page=1&pageSize=1"
# re3data: все репозитории одним ответом
curl -s "https://www.re3data.org/api/v1/repositories"
```

Ответ в RDF, SKOS или SDMX по адресу записи — хороший знак. Если каталог отвечает только страницей для людей, а программе — ошибкой или проверкой от роботов, на него трудно опереться. Сводный список — в чек-листах, общие команды — в главе об API и массовой выгрузке.

Чем отличается от соседей

Сосед	Главное отличие
Портал открытых данных	портал отдаёт файлы и таблицы, каталог метаданных — описания, словари и перечни; портал справочников — пограничный случай
Каталог показателей	реестр SDMX хранит справочники и структуры таблиц, а каталог показателей отдаёт сами значения
Агрегатор и поисковик данных	агрегатор собирает как можно больше наборов для поиска; каталог каталогов описывает сами каталоги и не подменяет поиск по всем данным
Каталог микроданных	каталог микроданных описывает обследования и выдаёт файлы по заявке; каталог описаний закрытых данных только сообщает, что данные есть и куда обращаться
Научный репозиторий	репозиторий хранит данные и выдаёт DOI; каталог метаданных хранит описания, а данные остаются у владельца
Корпоративный каталог данных	описывает таблицы и потоки внутри компании для её сотрудников; публичный каталог метаданных открыт всем и описывает общие словари и данные многих организаций

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: число каталогов, владельцы, страны, платформы с долей API, интерфейсы. Запросы — в приложении.
- SKOS, рекомендация W3C 2009 года; XKOS, DDI Alliance, версия 1.0; OWL 2; ISO/IEC 11179-3:2023, четвёртая редакция от 16 января 2023 года; SDMX
- Глобальный реестр SDMX: справочник `CL_FREQ` и число структур по API, 12 октября 2026 года; тема «Транспорт» в словарях ЕС; открытый API I14Y — проверено 12 октября 2026 года
- Акт об управлении данными, регламент 2022/868, статья 8; Европейский реестр охраняемых данных на data.europa.eu — число записей и стран по API поиска, 12 октября 2026 года; правила сбора для национальных точек
- Регламент о Европейском пространстве медицинских данных, краткое изложение на EUR-Lex — сроки; проект исполнительного акта по статье 77, 2026 год; HealthDCAT-AP Release 7, проект спецификации от мая 2026 года
- Спецификация FAIR Data Point, исходный код и выпуски, индекс точек — состояние точек по API индекса, 12 октября 2026 года
- Skosmos и его выпуски; OntoPortal; Ontology Lookup Service; MOLGENIS; Metadata Browser — проверено 12 октября 2026 года
- Fusion Metadata Registry и страница проекта BIS

- Aristotle Metadata Registry, открытый репозиторий — последнее изменение 28 августа 2020 года, в репозитории только объявление о переходе на облачный сервис, проверено 12 октября 2026 года
- re3data — число репозиториев по API; BARTOC — число словарей по API; Linked Open Vocabularies — счётчик на главной; OLS — число онтологий; всё 12 октября 2026 года; FAIRsharing
- Каталоги каталогов, проверено 12 октября 2026 года: dataportals.org — счётчик на главной; Open Data Inception и исходный набор — число записей по API; каталог экосистемы CKAN; dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года
- Public Service Data Catalogue, Ирландия — счётчик на главной, 12 октября 2026 года; UK Data Service, Access conditions — три уровня доступа, проверено 12 октября 2026 года
- Logius, Stelselcatalogus — назначение и ответственность держателей реестров; Stelselcatalogus обновлён, 03.06.2026
- schema.gov.it — национальный каталог семантики данных Италии, проверено 12 октября 2026 года; российские порталы НСИ из-за рубежа 12 октября 2026 года не открылись, данные о них — по постам
- И. Бегтин, Telegram: запуск datacatalogs.ru, 03.11.2021; datacatalogs.ru на Airtable, 263 каталога, 23.06.2022; Open Data Inception и datacatalogs.org, 15.12.2021; метакаталоги: Geolode, DataShades, Open Data Inception, 15.05.2023; FAIRsharing, 02.01.2023; открытый каталог неоткрытых данных Ирландии, 27.05.2022; UK Data Service и три режима доступа, 26.10.2023; порталы НСИ в России, 22.05.2023; schema.gov.it, 03.06.2023; портал справочников Минздрава, 14.06.2024

Поисковики и агрегаторы

Поисковик данных не публикует своих наборов. Он забирает описания из многих каталогов и даёт по ним один поиск. Сами файлы остаются у источника, человек переходит туда по ссылке.

Практический вывод для двух сторон. Тому, кто ищет данные, поисковик полезен как первый шаг, когда неизвестно, у кого они лежат. Но ответ «ничего не найдено» не значит, что данных нет: каждый поисковик видит только свою часть мира. Владельцу каталога поисковики дают пользователей почти даром, если каталог отдаёт описания стандартным способом. А собственный поисковик организации нужен редко: его дорого вести, и такие сервисы закрываются чаще других.

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) поисковиков 124. Это самый размытый тип в реестре: больше половины записей — витрины двух систем, а многие крупнейшие поисковики записаны другими типами. Разбор — в разделе «Число поисковиков — не мера».

Агрегатор или поисковик

Оба забирают чужие описания и сами ничего не публикуют. Разница в том, для кого сделаны страницы. У агрегатора у каждого набора есть своя страница с постоянным адресом, и её специально показывают обычным поисковикам вроде Google: есть карта сайта, страница открывается без скриптов. Агрегатор хочет, чтобы его находили. Поисковик показывает только выдачу на запрос, а отдельных страниц, которые индексировал бы кто-то ещё, у него нет: человек сразу уходит к источнику.

- **NIAID Discovery Portal** Национального института аллергии и инфекционных болезней США собирает описания биомедицинских наборов из десятков репозиториев. На 12 октября 2026 года в нём 14,2 миллиона записей, у каждой своя страница, а в `robots.txt` перечислены 26 карт сайта. Это агрегатор.
- **find.data.gov.scot** называет себя поисковиком по данным Шотландии, но у каждого из 26,5 тысячи наборов из 70 источников своя страница, и все они перечислены в карте сайта (октябрь 2026 года). По устройству это тоже агрегатор.
- **Google Dataset Search** своих страниц наборов не держит: карточка в выдаче ведёт на сайт источника. Это поисковик.

Для владельца данных разница практическая. Поисковик лишних копий не создаёт. Агрегатор — создаёт: в Google рядом с вашей страницей может оказаться страница агрегатора с тем же набором. Как сделать так, чтобы поисковик считал главной вашу, — в главе о находимости.

Дальше в главе оба вида называются поисковиками, если разница не важна.

Кто создаёт

Поисковики данных делает в основном наука. Исследовательские инфраструктуры и фонды платят за поиск по научным результатам, потому что требуют от учёных публиковать данные и хотят видеть, что опубликовано.

Владелец	Поисковиков	Доля
Университеты и научные организации	84	68%
Международные организации	17	14%
Компании	8	6%
Сообщества и общественные организации	7	6%
Центральные органы власти	6	5%
Прочие	2	2%

Научные сети — OpenAIRE и построенные на её графе витрины сообществ, EUDAT с поисковиком B2FIND, национальные поисковики исследовательских данных Швеции, Австралии, Китая, Турции.

Международные организации ищут по своей отрасли: поисковики Международной инициативы по прозрачности помощи (IATI) по проектам помощи развитию, eChemPortal ОЭСР по сведениям о химических веществах, Ocean InfoHub по данным об океане, GARDIAN исследовательского консорциума CGIAR по сельскохозяйственным данным.

Компании — Google с Dataset Search, GeoSeer с поиском по геосервисам, Transitland с каталогом расписаний транспорта.

Государства поисковики делают редко. Чаще они строят национальный портал, который сам собирает ведомства и регионы, — американский data.gov, европейский data.europa.eu. По смыслу это та же работа на уровень выше, но реестр записывает такие сайты порталами открытых данных. Заметное исключение — Шотландия, где поисковик появился раньше портала.

По территории: 37 поисковиков охватывают Евросоюз целиком, 21 — весь мир, 23 приходятся на Италию, 8 — на Германию. В остальных странах — один-два.

Число поисковиков — не мера

Из 124 записей отдельных сервисов со своим индексом около сорока. Остальное устроено так:

- **50 — витрины OpenAIRE CONNECT.** Это облачный сервис OpenAIRE: сообщество, университетский альянс или исследовательская инфраструктура получает свой сайт поиска, который показывает срез общего графа OpenAIRE по правилам этого сообщества. Своего сбора у витрины нет. По сути это один поисковик с пятьюдесятью входами.
- **28 — системы VIVO.** VIVO — свободная программа для профилей учёных организации: люди, публикации, проекты, иногда наборы данных. Из 28 установок 17 — итальянские UNIFIND, которые университетам Италии ставит один консорциум. Отсюда и 23 «поисковика» Италии.

- **6 — Idra.** Все шесть — демонстрационные и тестовые стенды, и все не действуют, хотя саму программу продолжают развивать.

С другой стороны, многие крупные поисковики в реестре записаны иначе. Dateno, NIAID Discovery Portal, Research Data Australia, GESIS Search и DataONE числятся научными репозиториями. B2FIND, Earthdata Search NASA, OmicsDI, data.europa.eu и data.gov — порталами открытых данных. Отчасти это честно: многие из них одновременно и хранят что-то своё, и собирают чужое. Но считать мир поисковиков по одному типу реестра нельзя. Есть и обратные ошибки: поисковик голландского архива DANS EASY записан действующим, хотя закрыт 2 декабря 2024 года, а его наборы переехали в новые хранилища DANS Data Stations.

Что собирают

Главная единица поисковика — запись о чужом наборе. В ней ровно то, что отдал источник: название, описание, издатель, ссылка на страницу набора, иногда лицензия, территория, даты и ссылки на файлы. Если источник не указал лицензию, у поисковика её тоже не будет: выдумывать он не имеет права.

Свою работу поисковик добавляет поверх:

- **приводит поля к одному виду** — страну, формат, лицензию, тему, чтобы по ним работали фильтры;
- **склеивает дубли** — один набор с городского, регионального и национального портала должен стать одной записью;
- **помнит источник и дату сбора**, чтобы было видно, откуда запись и насколько она свежая;
- **иногда смотрит в сами данные** — скачивает файл и разбирает колонки, как Auctus, или кладёт копию данных в своё хранилище.

Что именно попадает в поисковик, зависит от его источников. Научные поисковики собирают записи DataCite и репозиториев, государственные — порталы открытых данных, геопоиск — каталоги слоёв и геосервисы, агрегаторы статистики вроде DBnomics — сами временные ряды (в октябре 2026 года — 94 источника и около 47 тысяч наборов рядов). Общего поисковика, который одинаково хорошо видел бы все типы данных, нет.

Как устроены

Внешне поисковики похожи: строка поиска, фильтры, список карточек. Устроены они по-разному, и от устройства зависит, что в них есть, а чего не будет никогда. Чаще всего встречаются семь моделей.

Поиск по разметке страниц

Поисковик ни с кем не договаривается. Он обходит веб, как обычный поисковик, и забирает описания из разметки schema.org/Dataset на страницах наборов. Так работает Google Dataset Search: в феврале 2023 года — больше 45 миллионов наборов с 13 тысяч сайтов, новее

официальной цифры Google не публиковал. Охват огромный, но только по тем сайтам, которые поставили разметку, и проверить её поисковик не может. Геопорталы и каталоги показателей разметку ставят редко, поэтому их данных здесь почти нет.

Сборщик по протоколам каталогов

Поисковик ведёт список источников и для каждого типа платформы держит свой сборщик: API CKAN, файлы DCAT, OAI-PMH, CSW, интерфейс ArcGIS. Так устроены Dateno, B2FIND, BASE и OpenAIRE. Dateno 12 октября 2026 года показывал 22 миллиона наборов из 5,2 тысячи источников; в конце 2024 года у него были сборщики больше чем для 15 типов платформ. B2FIND консорциума EUDAT собирает научные репозитории Европы: в октябре 2026 года в нём 1,47 миллиона записей.

Такой модели нужен реестр источников с ручной проверкой и сборщик под каждую платформу. Самописный портал со своим интерфейсом она видит последним или не видит вовсе. У Dateno такой реестр — dataportals-registry, по которому посчитаны цифры этой книги. Реестр изначально создавался для поисковика, и Dateno индексирует порталы по их описаниям в реестре: по записанной платформе и адресам API выбирает сборщик.

Отраслевой агрегатор

Собирает источники одной отрасли и знает её схемы описаний, поэтому даёт фильтры, которых нет у общих поисковиков: организм и тип эксперимента, прибор и спутник, вещество и регламент. Примеры:

- **NIAID Discovery Portal** — биомедицина, 14,2 миллиона записей, 62% из одной базы NCBI GEO;
- **OmicsDI** — данные «омиксных» исследований (геномика, протеомика и подобные), 4,9 миллиона наборов из 29 баз (октябрь 2026 года);
- **DataONE** — науки о Земле и окружающей среде, около 1,08 миллиона текущих описаний без заменённых версий из 83 зарегистрированных репозиториев-участников;
- **Earthdata Search** — данные наблюдений Земли NASA, 45 тысяч коллекций; каждая коллекция — серия продуктов, в которой могут быть миллионы файлов;
- **eChemPortal** ОЭСР — сведения о химических веществах из государственных баз разных стран;
- **Transitland** — расписания общественного транспорта в формате GTFS со всего мира.

Отраслевой агрегатор — самый живучий вид: у него есть заказчик, который понимает, зачем он нужен.

Национальный поисковик

Страна или регион вместо одного общего портала собирает описания из сотен местных источников. find.data.gov.scot — 26,5 тысячи наборов из 70 источников; портал data.gov.scot появился позже, в 2026 году. Research Data Australia — поиск по австралийским университетам, научным организациям и ведомствам: в июле 2025 года около 224 тысяч записей, из них около 96 тысяч с данными, доступными онлайн. Похожие сервисы есть у Швеции (researchdata.se), Китая (Findata) и Турции (академический архив Harman).

Такой модели нужны договорённость с источниками, сбор по DCAT и ссылка на первоисточник в каждой карточке. Шотландский путь «сначала поиск, потом портал» разобран в главе «Федерации и агрегация».

Обходчик сервисов

Поисковик сам находит серверы определённого стандарта и опрашивает их, не спрашивая владельцев. GeoSeer обходит геосервисы WMS, WFS, WCS и WMTS: 3,5 миллиона слоёв с 4,2 тысячи серверов. STAC Index ведёт список каталогов спутниковых снимков и других пространственных файлов в стандарте STAC: в октябре 2026 года 149 каталогов, из них 63 — серверы с API и 86 — статические наборы файлов. Обходчик видит то, что не размечено и ни в какой каталог не внесено, но только в своём стандарте.

Витрина над общим индексом

Сама ничего не собирает, а показывает срез большого индекса по правилам сообщества: витрины OpenAIRE CONNECT, общий указатель точек FAIR Data Point, системы профилей VIVO. Запустить такую витрину дёшево, но содержимое у неё то же, что в общем индексе, со всеми его дырами.

Поиск по содержимому

Поисковик скачивает сами файлы, разбирает колонки, их типы, даты и координаты и ищет уже по ним. Так работает Auctus Нью-Йоркского университета: в нём можно найти набор, который дополняет вашу таблицу по общей колонке, например по коду района или дате. Цена — охват: в октябре 2026 года в Auctus около 55 тысяч наборов из 64 источников, половина — с Humanitarian Data Exchange. Скачать и разобрать каждый файл дорого, поэтому поиск по содержимому пока остаётся нишевым.

Что из этого следует

Модель	Пример	Без чего не обойтись	Слабое место
По разметке страниц	Google Dataset Search	разметка schema.org у источников	верит разметке на слово, не видит неразмеченное
Сборщик по протоколам	Dateno, B2FIND, OpenAIRE	реестр источников, сборщики под платформы, склейка дублей	самописные порталы без стандартного интерфейса
Отраслевой агрегатор	NIAID, DataONE, Earthdata	отраслевые схемы описаний и фильтры	видит только свою отрасль
Национальный поисковик	find.data.gov.scot	договорённость с источниками, сбор по DCAT	держится на одной команде и одном бюджете
Обходчик сервисов	GeoSeer, STAC Index	поиск серверов, разбор ответов стандартных сервисов	видит только свой стандарт
Витрина над индексом	OpenAIRE CONNECT	общий индекс, настройка правил отбора	содержимое не богаче общего индекса
По содержимому	Auctus	скачивание и разбор файлов	малый охват

Стандарты

Своих стандартов у поисковика почти нет: он читает то, что источник отдаёт машинно. Поэтому его стандарты — это стандарты источников:

- **schema.org/Dataset** на странице набора — так читает Google Dataset Search. Без разметки набор для него не существует.
- **DCAT и его профили** — так общеевропейский и национальные порталы забирают друг друга, а независимые поисковики — порталы открытых данных.
- **OAI-PMH** по руководствам OpenAIRE и **схема DataCite** — так научные агрегаторы забирают репозитории и записи с DOI.
- **CSW** — так собирают каталоги слоёв геопорталов.
- **STAC** — так перечисляют и связывают каталоги снимков.
- **FAIR Data Point** — так научные каталоги публикуют описания в RDF и регистрируются в общем указателе.

Дальше описания приводят к одному виду, убирают дубли и ранжируют. Как устроена такая цепочка, разобрано в главе «Федерации и агрегация».

Единого стандарта выдачи у поисковиков нет: API у каждого свой, у одних открытый (DataCite, OpenAIRE, data.europa.eu), у других платный (GeoSeer) или по ключу (Dateno), у Google Dataset Search его нет совсем.

Обязанности у сторон простые. Источник должен отдавать стабильный адрес выгрузки и указывать лицензию у набора. Поисковик не должен показывать описание, у которого исходная страница уже не открывается, или хотя бы должен это помечать.

Платформы

Платформа	Что это	Поисковиков в реестре	API
OpenAIRE CONNECT	облачный сервис витрин сообществ над графом OpenAIRE	50	общий API OpenAIRE
Своя разработка	Google Dataset Search, GeoSeer, find.data.gov.scot, STAC Index и другие	35	у 20%
VIVO	свободная система профилей учёных, лицензия BSD-3-Clause	28	у 25%
Idra	свободная платформа федерации порталов открытых данных, лицензия AGPL-3.0	6	все шесть — недействующие стенды
FAIR Data Point	указатели точек FAIR Data Point	2	у обоих

По одному поисковику работают на DataLad (указатель наборов, которые хранятся в git-репозиториях под управлением DataLad) и на программе европейской геофизической инфраструктуры EPOS.

Idra — самый близкий к «готовому поисковику» вариант. Это свободная программа из каталога FIWARE: она федерирует порталы на CKAN, DKAN и Socrata, брокеры данных Orion, принимает выгрузки DCAT-AP, умеет разбирать страницы порталов без API, приводит

описания к DCAT-AP и отдаёт их через REST API и SPARQL. Выпуск 4.0.0 вышел в марте 2025 года, изменения в код вносили и в октябре 2026 года. Но в реестре нет ни одной действующей установки: все шесть — демонстрационные стенды разработчиков и проектов, которые закрылись. Выбирая Idra, стоит заранее найти живую установку у похожей организации.

Платформа портала со сбором. Национальный или региональный агрегатор чаще строят на платформе портала, которая умеет собирать чужие каталоги:

Платформа	Сбор чужих каталогов	Что ещё нужно
Piveau	в ядре: сбор и проверка качества описаний DCAT — основа платформы, на ней работает data.europa.eu	—
CKAN	через расширения: каркас сбора ckanext-harvest, сбор RDF в ckanext-dcat, сбор CSW в ckanext-spatial	сборщики для других протоколов — отдельные расширения или своя доработка
Idra	в ядре: CKAN, DKAN, Socrata, выгрузки DCAT-AP, разбор страниц	своя проверка, что проект жив

На CKAN, например, работает B2FIND, причём на старой ветке 2.9.8, которая уже не получает обновлений, со своим расширением сбора. Реестр запишет такой сайт порталом открытых данных, хотя по сути он агрегатор. Подробности о расширениях — в главе о CKAN.

Своя разработка — сборщики плюс поисковый движок Elasticsearch, OpenSearch или Solr. Так сделаны почти все крупные поисковики. Готовой программы, которая из коробки собирает и CKAN, и DSpace, и GeoNetwork, и ArcGIS, нет, поэтому большой общий поисковик — всегда собственная разработка.

Примеры

Общие

- Google Dataset Search — ищет по разметке schema.org на любых сайтах.
- Dateno — сборщик порталов открытых данных, геопорталов, научных репозиториев и статистики, 22 миллиона наборов; API по ключу.

Научные

- OpenAIRE Explore — научные результаты Европы и мира, в том числе 108 миллионов записей о наборах данных.
- DataCite Commons — все записи с DOI DataCite, включая наборы данных.
- B2FIND — научные репозитории Европы, работает на CKAN.

Отраслевые

- NIAID Discovery Portal — биомедицина.
- OmicsDI — геномика, протеомика и подобные данные.
- DataONE — науки о Земле и окружающей среде.
- Earthdata Search — данные наблюдений Земли NASA.
- Transitland — расписания общественного транспорта.

Национальные и региональные

- find.data.gov.scot — Шотландия.
- Research Data Australia — научные данные Австралии.
- data.europa.eu — национальные порталы стран Европы; формально портал открытых данных, по работе — агрегатор агрегаторов.

Геоданные

- GeoSeer — слои геосервисов WMS, WFS, WCS и WMTS.
- STAC Index — каталоги снимков и других пространственных файлов.

По содержимому

- Auctus — ищет по колонкам и значениям внутри файлов.

Как оценить поисковик

Число наборов на главной странице говорит мало. Поисковик, в который залили два-три крупных научных агрегатора, получает десятки миллионов записей и остаётся поисковиком по научным данным. Полезнее проверить пять вещей.

1. **Разнообразие данных.** Находятся ли рядом открытые данные государства, геоданные, научные наборы, статистика, данные для машинного обучения. Если почти всё одного типа, это отраслевой поиск, даже если он называет себя общим.
2. **Охват по странам.** Находится ли что-нибудь по маленькой стране, а не только по США и Западной Европе. Простая проверка — запрос с названием небольшого государства на его языке.
3. **Источники.** Сколько каталогов собрано и опубликован ли их список. Если список закрыт, нельзя понять, чего в поиске нет.
4. **Ссылки на файлы.** Ведёт ли карточка к самому файлу или API, или только к странице набора у источника, а то и просто сообщает, что набор существует.
5. **Фильтры.** Можно ли сузить выдачу по стране, типу источника, формату, лицензии, теме. Ранжировать наборы по ссылкам почти нечем, поэтому без фильтров длинная выдача бесполезна.

Эти критерии тянут в разные стороны: гнаться за числом записей проще, чем за разнообразием и ссылками на файлы. Хороший поисковик честно показывает, на что он сделал ставку.

Сколько в крупных индексах

Цифры ниже нельзя сравнивать напрямую: каждый индекс считает «набором» своё. Они показывают порядок величин и то, из чего эти величины складываются.

Индекс	Что собирает	Записей «наборов»	Дата	Оговорка
Google Dataset Search	страницы с разметкой schema.org	более 45 млн с 13 тыс. сайтов	февраль 2023 года, по данным Google	новее официальной цифры нет

Индекс	Что собирает	Записей «наборов»	Дата	Оговорка
DataCite	наборы с идентификатором DOI	76,4 млн	12 октября 2026 года	42,8 млн — от одного японского института термоядерного синтеза; ссылка на содержимое указана примерно у 12 тыс. записей
OpenAIRE Graph	научные результаты, в том числе данные	108 млн	12 октября 2026 года	ссылка ведёт на страницу набора у источника, а не на файл
BASE	научные репозитории по OAI-PMH	25 млн, из них 22 млн из DataCite	ноябрь 2024 года, по посту И. Бегтина	в октябре 2026 года сайт закрыт от программ проверкой на робота, перепроверить не удалось
Dateno	порталы открытых данных, геопорталы, научные репозитории, статистика	22 млн из 5,2 тыс. источников	12 октября 2026 года, по счётчику на главной	в ноябре 2024 года было 19 млн из более чем 5 тыс. каталогов
NIAID Discovery Portal	биомедицинские репозитории	14,2 млн	12 октября 2026 года	62% записей — из одной базы, NCBI GEO
OmicsDI	базы данных геномики, протеомики и подобных исследований	4,9 млн из 29 баз	12 октября 2026 года	часть этих баз собирает и NIAID
GeoSeer	слои геосервисов WMS, WFS, WCS, WMTS	3,5 млн с 4,2 тыс. серверов	12 октября 2026 года	видит только сервисы этих стандартов
B2FIND	научные репозитории Европы	1,47 млн	12 октября 2026 года	—
DataONE	репозитории наук о Земле и окружающей среде	1,08 млн	12 октября 2026 года	без заменённых версий; со всеми версиями — 1,64 млн
Earthdata (CMR)	данные наблюдений Земли NASA и партнёров	45 тыс. коллекций	12 октября 2026 года	коллекция — серия продуктов, а не файл

Три оговорки важнее самих чисел. Первая — нарезка: больше половины «наборов» DataCite дал один институт, который присваивает DOI каждому эксперименту, а NASA, наоборот, считает целую серию файлов одной коллекцией. Почему поле «число наборов» у разных каталогов различается на порядки, разобрано в главе «Мир каталогов в цифрах». Вторая — повторы: одни индексы собирают другие, поэтому 22 из 25 миллионов записей BASE — это DataCite. Третья — ссылки на файлы. Научные индексы собирают описания по OAI-PMH и из DataCite, а там списка файлов обычно нет. Из 76,4 миллиона записей DataCite типа «набор данных» ссылка на содержимое (поле `contentUrl`) есть примерно у 12 тысяч. Карточка сообщает, что набор существует, но скачать его из поиска нельзя.

Поиск геоданных

Google Dataset Search геоданные почти не видит: большинство геопорталов не отдают разметку schema.org. Поэтому для них есть отдельные инструменты, и у каждого своё слепое пятно.

GeoSeer обходит серверы стандартных геосервисов. В октябре 2026 года в нём 3,5 миллиона слоёв с 4,2 тысячи серверов в 97 странах, больше всего — в доменной зоне Германии. Поиск на сайте бесплатный, API платный. Список источников не опубликован, а фильтров по стране

и теме, по наблюдению 2024 года, нет. Главное ограничение — GeoSeer видит только сервисы WMS, WFS, WCS и WMTS. Слой, который выложен на портале открытых данных файлом или отдаётся через ArcGIS REST, для него не существует. Поиск ArcGIS Hub, наоборот, видит только сайты на облачной платформе Esri.

Такие сервисы ещё и недолговечны. Финская компания Spatineo вела открытый каталог Spatineo Directory: в 2023 году — больше 87 тысяч адресов геосервисов, почти все из Европы и Северной Америки. К октябрю 2026 года каталог закрыт, на его месте — платный поиск для ArcGIS Pro. Как ищут геоданные внутри самих геопорталов, разобрано в главе о геопорталах.

Где ломается

Устаревшая копия. Поисковик забрал описание год назад, источник с тех пор переехал или убрал набор, а карточка в поиске осталась и ведёт на мёртвую ссылку.

Тройной учёт. Один набор попал в поиск трижды: с муниципального портала, с регионального и с национального. Агрегатор выглядит богаче источников и при этом завышает число наборов.

Агрегатор собирает агрегатор. Поисковику проще подключить один крупный агрегатор, чем сотню его источников. Но тогда он получает вторичные копии с уже потерянными полями и повторы: 22 из 25 миллионов записей BASE пришли из DataCite. Похожий выбор стоит перед агрегаторами статистики: брать ряд у национальной статслужбы или у Всемирного банка, который его уже перепечатал. Правило простое — собирать первоисточник, а агрегатор подключать только для тех источников, до которых самому не дотянуться.

Чужая копия вместо источника. В индексе Dateno в 2024 году к России относились 244 тысячи наборов, из них 195 тысяч давал геопортал одного научного центра. Это оказался клон описаний наборов метеослужбы США NOAA, а не российские данные. Агрегатору нужен реестр источников с ручной проверкой, иначе один такой сайт искажает всю статистику по стране.

Поиск только по описаниям. Найти набор, в котором есть колонка с кодом организации или с координатами, ни Google Dataset Search, ни DataCite Search не умеют: поля и значения внутри файлов они не индексируют. Auctus показывает, что это возможно, но пока на десятках тысяч наборов, а не на миллионах.

Разметке верят на слово. Всё, что Google Dataset Search знает о наборе, написал сам владелец страницы, в том числе отметку «доступно бесплатно». Проверить её поисковик не может. Разметку schema.org/Dataset ставят и продавцы данных ради продвижения, и их карточки перемешаны с открытыми наборами. В ноябре 2024 года на запрос о населении Андорры даже с фильтром «бесплатно» первыми шли страницы платных сервисов статистики, а данные ООН — где-то ниже. Тот же запрос по-испански не дал ничего, хотя данные об Андорре есть у статистических и геопорталов, которые разметку не ставят.

Ранжировать нечем. Ранжировать наборы по ссылкам, как веб-страницы, почти не получается: на набор редко ссылаются. Поэтому поиск по данным держится на фильтрах — страна, тип каталога, формат, лицензия, тема, — и качество фильтров зависит от того,

насколько аккуратно источники заполнили описания.

Поисковики закрываются. Сбор надо поддерживать годами: источники переезжают, меняют платформы и API. Spatineo Directory закрыт, все шесть установок Idra в реестре не действуют, поиск DANS EASY закрыт вместе с архивом в декабре 2024 года, BASE закрылся от программ проверкой на робота. Если работа опирается на поисковик, сохраняйте ссылку на первоисточник, а не на карточку поисковика.

В грубой оценке доверия, которую считает реестр, поисковикам снижают балл: они не хранят данные и не отвечают за лицензию источника. Это не приговор, а напоминание не путать поиск с первоисточником. Подробнее в главе о методиках.

Как искать данные через поисковики

Начинайте не с самого большого поисковика, а с того, который знает ваш тип данных.

Что ищете	Где искать сначала
Данные органов власти	национальный портал страны, в Европе — data.europa.eu; затем Dateno
Статистику	сайт статслужбы и её каталог показателей, DBnomics; затем Dateno
Геоданные	национальный геопортал, GeoSeer, Dateno, поиск ArcGIS; для снимков — STAC Index
Научные данные	DataCite Commons, OpenAIRE Explore, отраслевой агрегатор (NIAID, OmicsDI, DataONE, Earthdata)
Данные с любых сайтов, в том числе коммерческие	Google Dataset Search
Набор, который дополняет вашу таблицу	Auctus

Несколько привычек экономят время:

- **Пишите запрос на языке страны и словами источника.** Статслужба Испании называет население «población», а не «population». Андорра в ноябре 2024 года показала, что один и тот же поисковик на разных языках находит разное.
- **Сужайте фильтрами, а не листайте.** Страна, формат, лицензия, тип источника отсекают больше лишнего, чем уточнение слов в запросе.
- **Проверяйте у источника.** Лицензию, дату обновления и сам файл смотрите на странице набора у издателя: поисковик мог забрать описание давно и с потерями. Ссылайтесь тоже на источник, а не на карточку поисковика.
- **Не верьте пустой выдаче.** Если ничего не нашлось, зайдите на портал или в каталог показателей нужной страны напрямую. Многие геопорталы и статистические сайты не попадают ни в один общий поисковик.

У нескольких поисковиков открытый API, и проверить, что о теме знает каждый из них, можно одной командой. Ответы проверены 12 октября 2026 года:

```
# DataCite: записи с DOI типа «набор данных», число — в meta.total
curl -s -g "https://api.datacite.org/dois?query=Andorra&resource-type-
id=dataset&page[size]=1"
# OpenAIRE Graph: наборы данных, число — в header.numFound
```

```
curl -s "https://api.openaire.eu/graph/v1/researchProducts?
type=dataset&search=Andorra&pageSize=1"
# data.europa.eu: наборы европейских порталов, число — в result.count
curl -s "https://data.europa.eu/api/hub/search/search?q=Andorra&filter=dataset&limit=1"
# NIAID Discovery Portal: биомедицина, число — в total
curl -s "https://api.data.niaid.nih.gov/v1/query?q=Andorra&size=0"
# Earthdata (CMR): число коллекций — в заголовке CMR-Hits
curl -sI "https://cmr.earthdata.nasa.gov/search/collections.json?keyword=Andorra"
# Auctus: поиск по колонкам и значениям внутри файлов
curl -s -X POST "https://auctus.vida-nyu.org/api/v1/search" \
  -H 'Content-Type: application/json' -d '{"keywords": ["Andorra"]}'
```

На этот запрос DataCite нашёл около 9,5 тысячи наборов, OpenAIRE — около 9,2 тысячи, data.europa.eu — 585, NIAID — 101, Earthdata — 60 коллекций. Числа говорят не о том, где больше данных об Андорре, а о том, как каждый поисковик ищет: научные индексы находят слово в описании любого набора, где страна упомянута хотя бы вскользь, а data.europa.eu видит только европейские порталы.

Чтобы ваш собственный каталог находили поисковики, нужны выгрузка описаний, разметка и регистрация в профильных реестрах. Это разобрано в главе о находимости, а для публикации без портала — в главе «Как сделать публикацию находимой».

Когда свой поисковик нужен, а когда нет

Свой поисковик данных «по всему миру» организации почти никогда не нужен: такие сервисы уже есть, и даже они с трудом удерживают охват. Обычно хватает одного из более дешёвых вариантов:

- **попасть в существующие поисковики** — отдавать описания в DCAT или OAI-PMH, ставить разметку schema.org, зарегистрироваться в профильных реестрах;
- **завести витрину над общим индексом** — научному сообществу или альянсу университетов, у которых результаты уже есть в OpenAIRE, достаточно витрины OpenAIRE CONNECT;
- **собрать наборы на платформе портала со сбором** — региону или стране, где источники уже работают на CKAN и других типовых платформах.

Свой поисковик оправдан в двух случаях. Первый — отраслевой: общие поисковики не знают схем описаний отрасли и не дают нужных фильтров, а у отрасли есть заказчик, готовый платить годами, как у NIAID или DataONE. Второй — национальный, когда данные уже разбросаны по сотне сайтов и быстро свести их в один портал нельзя, как в Шотландии. В обоих случаях до старта стоит ответить на три вопроса: кто ведёт реестр источников, кто чинит сборщики, когда источники меняют платформу, и что будет с адресами карточек, если проект закроется. Как записать такое решение с обоснованием — в главе «Как обосновать решение».

Чем отличается от соседей

Сосед	Главное отличие
Портал открытых данных	публикует свои наборы и отвечает за них. Если портал и принимает публикации, и собирает чужие каталоги, как data.gouv.fr, это гибрид

Сосед	Главное отличие
Научный репозиторий	хранит файлы и присваивает им DOI. Поисковик ничего не хранит и DOI не выдаёт
Каталог метаданных	описывает каталоги, структуры и справочники. Каталог каталогов вроде re3data отвечает на вопрос «где искать», поисковик — «что есть»
Геопортал	показывает слои на карте и отдаёт сервисы. GeoSeer ищет по слоям, но сам сервисов не держит
Каталог показателей	хранит сами ряды и их методику. Агрегатор статистики вроде DBnomics ближе к каталогу показателей, потому что копирует ряды, а не только описания
Копия чужих карточек	перепечатывает описания без ссылки на источник и без даты сбора. Это не агрегатор, а копия

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: владельцы, территории, платформы и доля API поисковиков, список записей, а также крупные поисковики, записанные другими типами. Запросы — в приложении.
- Разметка Dataset в документации Google, руководства OpenAIRE, STAC, FAIR Data Point
- OpenAIRE CONNECT — описание сервиса витрин сообществ; OpenAIRE Graph API — запрос `researchProducts?type=dataset`, 12 октября 2026 года
- VIVO и репозиторий VIVO — лицензия BSD-3-Clause, выпуск 1.15.1 от 25 июня 2026 года
- Idra — описание возможностей в README, лицензия AGPL-3.0, выпуск 4.0.0 от 27 марта 2025 года, последние изменения — 8 октября 2026 года; Piveau
- B2FIND — версия платформы по ответу status_show и число записей через API CKAN, 12 октября 2026 года
- Dateno — счётчики наборов, источников и стран на главной странице, 12 октября 2026 года
- NIAID Discovery Portal — число записей и распределение по источникам: запрос к API портала, карты сайта в robots.txt, 12 октября 2026 года
- OmicsDI — число наборов через API поиска и список баз через API баз, 12 октября 2026 года
- DataONE — число описаний через поисковый API и список репозиториев-участников через реестр узлов, 12 октября 2026 года
- Common Metadata Repository NASA — число коллекций в заголовке ответа `CMR-Hits`, 12 октября 2026 года
- STAC Index — список каталогов через API, 12 октября 2026 года
- Auctus — число наборов по источникам через API статистики, исходный код, 12 октября 2026 года
- DBnomics — число источников и наборов через API, 12 октября 2026 года
- find.data.gov.scot — счётчики на главной и карта сайта, 12 октября 2026 года
- DANS EASY — сообщение о закрытии 2 декабря 2024 года и переносе наборов в DANS Data Stations, проверено 12 октября 2026 года

- Datasets at your fingertips in Google Search, блог Google Research, 28 февраля 2023 года — 45 млн наборов с 13 тыс. сайтов
- REST API DataCite — запросы к `api.datacite.org/dois` с фильтром по типу «набор данных», по издателю и по наличию `contentUrl`, 12 октября 2026 года
- Статистика GeoSeer и условия API, 12 октября 2026 года
- Ответы API на запрос «Andorra» в DataCite, OpenAIRE, data.europa.eu, NIAID и CMR — 12 октября 2026 года
- И. Бегтин, Поисковые системы по данным как глобальные системы обнаружения данных, май 2022 — поиск только по описаниям
- И. Бегтин, Открытые данные в мире, в фактах и тезисах, март 2023 — продавцы данных в выдаче Google Dataset Search
- И. Бегтин, Обновления в Dateno, ноябрь 2024 — фильтры вместо ранжирования по ссылкам; Почему данные по России так непросто найти, ноябрь 2024 — клон описаний NOAA
- И. Бегтин, Telegram: критерии глобального поисковика по данным, 30.06.2023; показатели наполнения индекса, 22.03.2024; агрегатор и поисковик на примере DataMed, 14.09.2024; NIAID Discovery Portal, 28.11.2024; find.data.gov.scot, 03.11.2024; обзор поисковиков с цифрами, 25.11.2024; DataCite Commons без ссылок на файлы, 25.09.2024; GeoSeer, 17.08.2024; Spatineo Directory, 30.07.2023; запрос об Андорре в Google Dataset Search, 16.11.2024; Auctus и поиск по содержимому, 02.06.2022; B2FIND, 11.08.2022; DBnomics и выбор между первоисточником и вторичным источником, 19.11.2024; сборщики Dateno для разных платформ, 31.12.2024; Research Data Australia, 17.07.2025
- Федерации и агрегация

Каталоги API, маркетплейсы и списки наборов

Три типа из этой главы встречаются редко. По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) это 115 каталогов API, 56 маркетплейсов и 107 списков наборов: вместе меньше 1% из 43,6 тысячи каталогов. Их легко не заметить в общей статистике и легко перепутать с «настоящим» порталом.

Каждый из них отвечает на свой вопрос о границе понятия. Что, если страницу в каталоге получает не набор данных, а программный интерфейс? Что, если набор — товар, и доступ к нему продают по договору? Что, если каталога нет вовсе, а есть страница со ссылками?

Главный вывод главы: все три типа хорошо работают как дополнение к каталогу данных и плохо — вместо него. Каталог API без карточек наборов прячет, какие данные стоят за интерфейсами. Маркетплейс с закрытыми описаниями прячет сам факт, что данные существуют. Список без лицензий и дат быстро становится тупиком. Лекарство во всех трёх случаях одно: открытое описание по общему стандарту рядом с основной витриной.

	Каталог API	Маркетплейс	Список наборов
Что получает свою страницу	программный интерфейс	товар: файл, подписка, доступ к таблице	ссылка на файл или чужой каталог
Каталогов в реестре	115	56	107
Главный владелец	центральные органы власти, 45%	компании, 70%	никто не преобладает
Своя разработка	74%	98%	89%
Программный доступ к самому каталогу	51%	21%	14%
Ограниченный доступ хотя бы к части данных	14%	91%	1%
Стандарт описания	OpenAPI, AsyncAPI, сервис данных DCAT	общего нет; ODRL и Dataspace Protocol в пространствах данных	чаще никакого; schema.org, Data Package

Строка «программный доступ» говорит о самом каталоге: можно ли программой получить список интерфейсов, товаров или ссылок. У каталога API она не про интерфейсы, которые он перечисляет.

Каталоги API

Что это. Портал для разработчиков: список программных интерфейсов органа власти, транспортной компании или банка с документацией, примерами запросов, а часто и с выдачей ключей и тестовой консолью. Это не то же самое, что API самого портала открытых данных из главы об API и массовой выгрузке. Там API — дверь в каталог. Здесь интерфейс и есть то, что каталог описывает.

У хорошей карточки интерфейса те же вопросы, что у карточки набора, плюс свои: что он отдаёт, базовый адрес, машиночитаемое описание, способ доступа (открыт, по ключу, по заявке), лимиты, лицензия данных, состояние сервиса, контакт и — самое частое упущение —

какие наборы данных за ним стоят.

Кто создаёт

Почти половину каталогов API ведут центральные органы власти (45%), ещё 23% — регионы, по 10–11% — компании и муниципалитеты. Больше всего записей у США (24), Финляндии (8), Великобритании (6), Швеции и Норвегии (по 5).

Около трети каталогов API в реестре — 35 из 115 — порталы разработчиков транспорта: метро и автобусных компаний США (MBTA, BART, WMATA, CTA), скандинавских операторов (Entur, Trafiklab, Digitraffic), Лондона, Новой Зеландии, Тайваня. Причина практическая: приложению с расписанием нужны данные о задержках в реальном времени, и транспортная компания выдаёт для этого ключ. Расписания при этом обычно отдают файлом GTFS, а движение машин — потоком GTFS Realtime.

Какие бывают

- **Национальный каталог государственных API.** api.gov.uk в Великобритании, API Setu в Индии, каталог gov.br/conecta в Бразилии, api.gov.au в Австралии, портал разработчиков Сингапура, каталоги Италии, Литвы, Японии, реестр API Нидерландов. Их часто делают отдельно от портала открытых данных: доступ к API почти всегда требует ключа, тестовой консоли и документации, а аудитория — разработчики, а не аналитики.
- **Портал разработчика одной организации.** Центральные банки (Риксбанк, Банк Финляндии, Банк Негара Малайзия), статистические службы (Stats NZ, польский GUS), метеослужбы (британский Met Office, индийский IMD), транспорт.
- **Раздел API в портале открытых данных.** Во Франции отдельный каталог api.gouv.fr влили в data.gouv.fr: первые карточки переехали 20 ноября 2024 года, в 2025 году старый сайт стали выводить из работы, а теперь его адрес ведёт в раздел сервисов данных портала. 12 октября 2026 года там 1265 интерфейсов: 1119 открытых, 37 с регистрацией и 109 с ограниченным доступом. У 1162 есть ссылка на машиночитаемое описание, у 1038 — на наборы, которые интерфейс отдаёт. Так же устроен каталог API Норвегии в общем национальном каталоге: 1071 сервис данных рядом с наборами. Финский каталог интерфейсов Suomi.fi сделан на CKAN.
- **Реестр интерфейсов государственных систем.** Эстонские RIHA и X-tee, лаосский GDX. Это учёт того, какие системы чем обмениваются, в основном для самих ведомств. Публике из него видно описание, но не доступ.
- **Общественные сборники.** APIs.guru собирает описания OpenAPI публичных интерфейсов в открытый репозиторий: 12 октября 2026 года — 2529 API и 3992 описания. Рядом publicapis.dev, коммерческие Postman Public API Network и RapidAPI Hub. Немецкий bund.dev — проект волонтёров, которые находят недокументированные государственные API Германии и пишут для них документацию за государство.
- **Реестры серверов MCP.** Официальный реестр MCP работает с 8 сентября 2025 года в предварительном режиме. Задача та же — найти интерфейс, — только читает такой каталог не программист, а программа-ассистент. Подробнее о MCP у каталогов данных — в главе о трендах.

Стандарты

OpenAPI описывает обычный интерфейс «запрос — ответ»: какие адреса, какие параметры, что придёт в ответ. По описанию инструменты сами строят страницу документации, тестовую консоль и клиентские библиотеки. Последняя версия — 3.2.0 от 19 сентября 2025 года. **AsyncAPI** делает то же для потоков событий: подписка на сообщения о задержках, показания датчиков. Версия 3.1.0 вышла 31 января 2026 года.

Сервис данных в DCAT. С версии DCAT 2 (2020) у словаря каталогов есть отдельная сущность для интерфейса: адрес, ссылка на описание вроде OpenAPI и список наборов, которые сервис отдаёт. Это мост между каталогом API и каталогом данных: одна запись говорит «набор такой-то можно получить вот этим интерфейсом». На этом мосту построены разделы API во французском и норвежском национальных каталогах.

Адрес для программ. RFC 9727 (июнь 2025 года) задаёт стандартный адрес `/.well-known/api-catalog`. По нему программа получает список интерфейсов сайта в формате linkset, не зная заранее, где лежит портал разработчика. Стандарт новый. 12 октября 2026 года ни один из одиннадцати проверенных национальных и крупных каталогов — британский, австралийский, нидерландский, итальянский, сингапурский, норвежский, французский, api.data.gov, Digitraffic, Entur, bund.dev — по этому адресу список не отдавал. Где код ответа был 200, приходила обычная страница сайта.

Отраслевые стандарты. В транспорте это GTFS и GTFS Realtime. В «умных городах» встречается NGSI-LD — интерфейс для данных датчиков в реальном времени, на нём работает индийская IUDX. У многих стран есть свои правила проектирования государственных API; пример — API Design Rules Нидерландов.

Платформы

Три четверти каталогов API — свои разработки. Из готовых решений чаще всего встречается портал разработчика при шлюзе API. Шлюз — программа, которая стоит перед интерфейсами организации: проверяет ключи, считает запросы, ограничивает нагрузку. Портал разработчика у него входит в комплект, поэтому каталог появляется почти «бесплатно».

- **Azure API Management** — 22 каталога в реестре, от Банка Финляндии и Fingrid до портала API Нью-Йорка и транспорта Вашингтона.
- **WSO2 API Manager** (свободный, Apache 2.0) — Met Office, латвийская VDAA. **Gravitee** — портал Французской Полинезии.
- **Статический сайт из файлов.** api.gov.uk собирается из одной таблицы CSV в открытом репозитории GitHub. APIs.guru — тоже репозиторий описаний OpenAPI, из которого генерируется сайт.
- **Платформы каталогов данных.** У udata, на которой работает data.gouv.fr, сервисы данных — отдельная сущность в ядре: в карточке есть поля для описания OpenAPI, лимитов, доступности, адреса заявки на доступ и связанных наборов. У CKAN отдельной сущности нет: расширение ckanext-dcat позволяет записать сервис доступа внутри ресурса набора и выгрузить его в DCAT.

Возможность	Портал шлюза API	Статический сайт из файлов	udata	CKAN
Страница интерфейса с документацией	ядро	генератор сайта	ядро, ссылка на описание OpenAPI	нет; ссылка из ресурса
Выдача ключей и учёт лимитов	ядро	нет	нет; есть поля «лимиты» и «как получить доступ»	нет
Связь интерфейса с наборами	нет	вручную	ядро	расширение ckanext-dcat: сервис доступа у ресурса
Видимость для поисковиков	слабая, см. «Где ломается»	хорошая	хорошая	хорошая

Выбор следует из задачи. Если организация выдаёт ключи к своим интерфейсам, портал шлюза уже есть, и вопрос только в том, чтобы описания были видны снаружи. Если нужно перечислить чужие интерфейсы, хватит статического сайта из таблицы. Если за интерфейсами стоят наборы данных, их стоит описать в каталоге данных и указать интерфейс как способ получения — по модели Франции и Норвегии.

Примеры

- Раздел API на data.gouv.fr — национальный каталог API внутри портала открытых данных, с типом доступа и связанными наборами в каждой карточке.
- api.gov.uk — каталог API британского государственного сектора, собранный из таблицы в открытом репозитории.
- API Setu — индийский каталог интерфейсов всех уровней власти.
- Digitraffic и Trafiklab — транспортные данные Финляндии и Швеции: интерфейсы реального времени и файлы расписаний.
- APIs.guru — открытый сборник описаний OpenAPI.
- bund.dev — документация к недокументированным государственным API Германии, сделанная волонтёрами.

Где ломается

Описание отстаёт от интерфейса. В документации одно поле, в ответе другое. Масштаб видно по APIs.guru: 12 октября 2026 года из 3992 собранных описаний 688 не проходили проверку формата, а 166 были недоступны по исходному адресу.

Каталог стареет. В британском каталоге 233 записи, и у 70 из них последнее обновление — 2020 год. Лицензия на данные указана у 36 (12 октября 2026 года). Каталог, который собрали один раз, перестаёт отражать живые интерфейсы так же, как заброшенный портал перестаёт отражать данные.

Описан интерфейс, но не доступ. Каталог перечисляет API, для которых нужен ключ, но не говорит, как его получить, кому дают и сколько ждать. В апреле 2024 года индийская IUDX описывала 189 наборов из 47 городов, в основном потоки реального времени, и ни один не был открыт: по сути это портал разработчика для партнёров, а не каталог открытых данных.

Поисковик не видит каталог. Портал разработчика при шлюзе часто собирает страницу скриптом в браузере. 12 октября 2026 года в HTML страницы со списком API у Банка Финляндии на Azure API Management не было ни одного названия интерфейса — только меню и кнопка входа. Такой каталог не находят ни поисковики, ни веб-архив. Карта сайта отмечена в реестре лишь у 26 каталогов API из 115.

Счёт интерфейсов раздувают. Из 1265 «API» на data.gouv.fr около 800 — автоматические интерфейсы к отдельным наборам с региональных порталов на платформе Data Fair, по одному на каждую таблицу. Это не ошибка: такой интерфейс полезен. Но число API в каталоге, как и число наборов, говорит об устройстве, а не о том, сколько у государства сервисов для разработчиков.

API есть, а каталога нет. Частый случай — интерфейсы существуют, но нигде не перечислены и не описаны. Они ломаются при каждом обновлении сайта, и их не сохраняет веб-архив. В России нет национального каталога государственных API: межведомственный обмен идёт через закрытые системы, а публичные интерфейсы госсайтов в 2021 году собирала «Инфокультура» — больше сотни. Московский apistore.mos.ru в 2021 году содержал 15 прототипов API и не обновлялся с 2019 года; в октябре 2026 года он не открывается из-за рубежа. В Германии ту же работу делают волонтёры bund.dev, во Франции — государство.

Интерфейс вместо выгрузки. Если данные можно получить только запросами по одной записи, тому, кому нужна вся база, приходится выкачивать её неделями и упираться в лимиты. Интерфейс — хорошее дополнение к массовой выгрузке, а не замена; подробнее — в разделе «Массовая выгрузка — не роскошь».

Как проверить каталог API

1. **Найдите машиночитаемое описание.** Есть ли у интерфейса файл OpenAPI, и совпадает ли он с настоящим ответом на пробный запрос?
2. **Найдите порядок доступа.** Сказано ли, нужен ли ключ, кто его выдаёт, какие лимиты и какая лицензия у данных?
3. **Найдите набор.** Есть ли у данных, которые отдаёт интерфейс, карточка в каталоге данных со ссылкой на интерфейс?
4. **Посмотрите как программа.** Видны ли названия интерфейсов в HTML страницы и есть ли список по адресу RFC 9727?

```
# RFC 9727: список интерфейсов сайта в формате linkset
curl -s -H "Accept: application/linkset+json" "https://example.org/.well-known/api-catalog"
# data.gouv.fr: сколько интерфейсов открыто; access_type=restricted — сколько закрыто
curl -s "https://www.data.gouv.fr/api/1/dataservices/?page_size=1&access_type=open"
# Норвегия: сервисы данных в национальном каталоге
curl -s -X POST -H "Content-Type: application/json" -d '{}' \
  "https://search.api.fellesdatakatalog.digdir.no/search/data-services"
# британский каталог API — одна таблица
curl -s "https://raw.githubusercontent.com/alphagov/api-catalogue/main/data/catalogue.csv"
# APIs.guru: сколько описаний собрано, сколько недоступно и невалидно
curl -s "https://api.apis.guru/v2/metrics.json"
```

Когда нужен

Каталог API нужен, когда у организации много интерфейсов и разработчику нужен один вход: где взять ключ, где документация, что изменилось. Одному API ведомства хватит страницы документации с описанием OpenAPI. Если за интерфейсом стоит набор данных, набор стоит описать и в обычном каталоге — со ссылкой на API как на способ получения. Тогда его найдёт и аналитик, который не ищет «API», и сборщик вроде data.europa.eu, который читает DCAT.

Маркетплейсы данных

Что это. Площадки, где данные продают или отдают по договору. Товаром может быть файл, подписка на обновления, доступ к интерфейсу или к таблице в облачном хранилище покупателя. От портала открытых данных маркетплейс отличает не оплата сама по себе, а договор на каждую выдачу: кто получил, на каких условиях, что ему можно делать с данными.

Кто создаёт

70% маркетплейсов в реестре — у компаний, 21% — у центральных органов власти. Доступ хотя бы к части данных ограничен у 91%, программный доступ к самой площадке отмечен у 21%. По странам первые — Южная Корея (16), США (12) и Китай (6), ещё 10 площадок работают на весь мир.

Цифры по странам показывают, что маркетплейс данных во многих странах — государственный проект, а не только бизнес:

- **Южная Корея.** Большая часть корейских записей — отраслевые платформы больших данных, которые Министерство науки и ИКТ запустило в 2019 году: десять платформ — финансы, окружающая среда, культура, транспорт, онкология, торговля, связь, малый бизнес, региональная экономика, лес. В 2020 году к ним добавили ещё шесть, в том числе пожарную безопасность и морское рыболовство. Все они собраны в общую карту данных: к концу 2021 года там было 6842 вида данных. Платформы смешивают бесплатные и платные данные.
- **Китай.** Государственные биржи данных: первая открылась в Гуйяне в 2015 году, Шанхайская биржа — 25 ноября 2021 года, с 20 продуктами в первый день. По обзору августа 2025 года действовали 49 таких площадок с 3944 поставщиками данных, а ещё 24 были ликвидированы или не открывались.
- **Япония.** JDEX и KYOTO DATA MARKETPLACE.
- **США.** Облачные площадки (AWS, Snowflake, Databricks, Google, Microsoft) и биржи, которые продают собственные рыночные данные: CME DataMine, Cboe DataShop, Nasdaq Data Link. Так же поступают корейская KRX и гонконгская HKEX.
- **Данные для ИИ.** Defined.ai, Nexdata, Centific продают наборы для обучения моделей: записи речи, размеченные изображения, тексты.
- **Пространства данных Европы.** Каталог Pontus-X в экосистеме Gaia-X, каталог Language Data Space, нидерландский DMI. О пространствах данных — ниже и в главе о трендах.

Какие бывают

Под одним словом «маркетплейс» живут четыре разные модели.

- **Государственные и отраслевые площадки.** Крайняя точка шкалы — data.europa.eu: общественная инфраструктура без платы и без сделок, то есть не маркетплейс, а то, с чем его сравнивают. Шанхайская биржа данных проводит платные сделки между компаниями. Catena-X в автопроме совмещает некоммерческое управление с платными услугами подключения. Британский Rail Data Marketplace, открытый в ноябре 2023 года, ведёт Rail Delivery Group: за первый год там появилось больше 140 продуктов, 1200 организаций и 4000 соглашений об обмене, а в октябре 2026 года — 269 продуктов, бесплатных и платных. Площадка не столько продаёт, сколько ведёт соглашения об использовании: при запуске организаторы обещали, что бесплатных данных будет большинство. Государство или отрасль здесь ещё и продвигает площадку, и может быть регулятором. Поэтому на ней появляются данные, которых нет больше нигде.
- **Данные внутри профессионального продукта.** Bloomberg и LSEG продают не наборы, а связку из данных, идентификаторов, лицензий и рабочего места трейдера. Это скорее способ встроить данные в продукт, чем отдельный рынок.
- **Площадки облачных провайдеров.** AWS, Snowflake, Google, Databricks. Они различаются тем, берут ли оплату сами и как доставляют данные. Общая логика одна: площадка увеличивает потребление основной облачной платформы.
- **Витрины поставщиков.** Datarade (на главной странице в октябре 2026 года — больше 2000 поставщиков) помогает найти поставщика и передать ему запрос, а договор, оплата и доставка остаются между сторонами. Это каталог с привлечением покупателей, а не биржа.

Стандарты

Описание товара. Общего стандарта нет, каждая площадка ведёт свою карточку. Ближе всего к общему языку описание дата-продукта: владелец, условия, способы доставки, тарифы. Шаблоны вроде ODPS разобраны в главе «Единицы публикации и форматы».

Условия использования. Там, где доступ договаривают между организациями, условия описывают машиночитаемо: ODRL — язык W3C для правил вида «можно использовать для исследований, нельзя передавать третьим лицам».

Протокол пространства данных. Dataspace Protocol (версия 2025-1, утверждена Eclipse Foundation 23 июля 2025 года) соединяет эти части. Участник публикует каталог своих наборов в DCAT, к каждому прикладывает условия в ODRL, а дальше протокол описывает переговоры о договоре и передачу данных. Свободная реализация — коннектор Eclipse Dataspace Components, на нём построен обмен в Catena-X. Важная деталь: описание набора здесь — обычный DCAT. Значит, его можно показать и снаружи, даже если сами данные закрыты.

Доставка без скачивания. Облачные площадки часто не отдают файл, а открывают покупателю доступ к таблице прямо в хранилище поставщика. AWS Data Exchange знает пять видов поставки: файлы, интерфейсы, доступ к таблицам Redshift, прямой доступ к

хранилищу S3 и права в Lake Formation. Databricks Marketplace работает на открытом протоколе Delta Sharing, который 10 июня 2026 года превратился в OpenSharing и перешёл в Linux Foundation.

Платформы

В реестре 55 маркетплейсов из 56 — собственные разработки, и это честная картина: тиражируемого ПО маркетплейсов почти нет. Облачные площадки — часть облака, у бирж данных свои системы. Открытые платформы каталогов — CKAN, Dataverse, InvenioRDM — оплату не принимают. Для пространств данных есть свободный коннектор EDC, но это канал обмена между участниками, а не витрина с оплатой.

Граница с порталами открытых данных размывается. 30 сентября 2025 года OpenDataSoft переименовалась в Huwise и стала продавать ту же платформу как маркетплейс дата-продуктов, в том числе для обмена внутри организации.

Почему площадки редко взлетают

Владельцам действительно востребованных данных посредник не нужен: клиенты и так приходят к ним на их условиях. Владельцы боятся, что посредник использует или перепродаст данные. А польза от «площадки» у данных меньше, чем у товаров: данные покупают редко, часто это разовая покупка или долгая подписка, и договариваться удобнее напрямую с поставщиком.

Работающие площадки обычно держит главный владелец данных или тот, у кого уже есть покупатели. Биржи продают собственные котировки. Rail Data Marketplace ведёт отраслевая организация железных дорог. Облачные площадки продают туда, где покупатель уже хранит данные. Независимой витрине чужих данных приходится годами тратиться на продвижение.

Примеры

- Rail Data Marketplace — отраслевая площадка железных дорог Великобритании: бесплатные и платные продукты, соглашения об использовании ведёт сама площадка.
- AWS Data Exchange, Snowflake Marketplace, Databricks Marketplace — облачные площадки с доставкой в хранилище покупателя.
- Nasdaq Data Link — финансовые и альтернативные данные биржи и сторонних поставщиков.
- Datarade — витрина поставщиков.
- JDEX и KYOTO DATA MARKETPLACE — японские площадки.
- Catena-X — пространство данных автопрома на Dataspace Protocol.

Где ломается

Описание закрыто до регистрации. Покупатель не понимает, что покупает, а поисковик не видит, что данные существуют. Рабочая схема другая: описания открыты, оплата и выдача идут отдельной системой.

Площадку не видно из-за рубежа. 12 октября 2026 года корейские платформы больших данных, их общая карта данных и китайские биржи из-за рубежа не открывались. Для внутреннего рынка это может быть приемлемо, но для иностранного покупателя, агрегатора и веб-архива таких площадок нет.

«Маркетплейс» как вывеска. Корейские платформы и Rail Data Marketplace во многом раздают данные бесплатно, а Huwise называет маркетплейсом платформу порталов открытых данных. Название не говорит, продают ли там что-нибудь. Тип определяют по тому, что получает пользователь: файл без условий — портал открытых данных; договор на каждую выдачу — маркетплейс.

Площадки закрываются. Из китайских бирж данных к августу 2025 года 24 были ликвидированы или не работали. data.world начинался как открытый облачный каталог, а 13 июля 2026 года закрыл открытое сообщество — история разобрана в главе об облачных хранилищах и хабах.

После подписки ничего не остаётся. При поставке «без скачивания» доступ к таблице исчезает в день окончания подписки. У файлов, которые покупатель успел выгрузить, судьбу решает договор: в AWS Data Exchange выгруженные файлы остаются у подписчика, но договор поставщика может требовать их удалить. Это стоит прочитать до покупки, а не после.

Как проверить маркетплейс

1. **Откройте карточку без входа.** Видны ли описание, состав полей, период, частота обновления, образец данных?
2. **Найдите условия.** Записаны ли условия использования в карточке, а не только в договоре, который покажут после оплаты?
3. **Узнайте способ доставки.** Файл, интерфейс или доступ к таблице в облаке — и сможете ли вы перенести данные к себе?
4. **Спросите, что останется.** Что можно хранить после окончания подписки, и можно ли ссылаться на данные в публикациях?

Когда нужен

Маркетплейс нужен, когда данные действительно продают или выдают по договору и на каждую выдачу нужны условия. Если цель — открыть данные, маркетплейс лишний: достаточно портала открытых данных или публикации без портала. Свою площадку для чужих данных строить почти никогда не стоит; почему — в карточке решения. Если данные продаются через облачную площадку, их описание стоит держать открытым и в обычном каталоге: тогда покупатель найдёт их поиском, а не только внутри облака.

Списки наборов

Что это. Страница, таблица или репозиторий со ссылками на файлы или на чужие каталоги. У списка нет поиска, фильтров и карточек: строка списка — ссылка и одна фраза описания. В реестре 107 таких списков, программный доступ — примерно у 14%.

Кто создаёт

Здесь нет главного владельца: 20% списков у компаний, 19% у центральных органов власти, 14% у муниципалитетов, 13% у сообществ, 12% у университетов. 89% — свои разработки, 16 лежат на GitHub или GitLab. Первые по странам — США (20) и международные проекты (12). Пять записей отнесены к Сейшельским островам: это криптобиржи, зарегистрированные там, которые выкладывают историю торгов.

Какие бывают

- **Тематическая подборка сообщества.** awesome-public-datasets на GitHub — около 79 тысяч звёзд в октябре 2026 года. Рядом списки по странам: awesome-nigeria-datasets, open-data-turkey. Последний в 2023 году оставался самым полным перечнем порталов Турции: национального портала там нет, а городские порталы и ГИС разбросаны.
- **Страница «Открытые данные» на сайте ведомства или города.** Кардифф, Галле, Арнсберг, Лодзь, Эр-Рияд, налоговая служба Саудовской Аравии. Файлы лежат на страницах обычной CMS — TYPO3, Drupal, SharePoint, — без отдельного портала.
- **Страница выгрузки одного издателя.** Поездки такси Нью-Йорка — по файлу Parquet на месяц и словари полей. Дампы Википедии, выгрузки OpenStreetMap по регионам, файлы расписаний транспорта. 13 записей из 107 — история торгов криптобирж. Например, data.binance.vision — листинг облачного хранилища, где рядом с каждым архивом лежит файл контрольной суммы.
- **Список, собранный из файлов описаний.** Registry of Open Data on AWS — сайт, который генерируется из 1238 файлов YAML в репозитории GitHub (12 октября 2026 года). В каждом файле одни и те же поля: название, описание, лицензия, частота обновления, контакт, адрес хранилища. Сам awesome-public-datasets устроен так же: README собирается автоматически из файлов YAML в соседнем репозитории apd-core.
- **Личные указатели на GitHub Pages.** Каталоги клинических данных, геоданных Индии, открытых данных Гонконга, которые ведёт один человек. Полезны как стартовая точка, но у них нет владельца, который ответит за полноту.

Стандарты

Чаще никаких. Довести список до машиночитаемого вида помогают:

- разметка schema.org — DataCatalog для страницы и Dataset для каждой строки;
- файл `datapackage.json` рядом с данными — его шаблон есть среди шаблонов книги;
- сам список в CSV, JSON или YAML в репозитории, из которого генерируется страница, — как у реестра AWS;
- файлы контрольных сумм рядом с архивами, как у Binance;
- файл DCAT со списком наборов — так статический сайт становится каталогом, который читают сборщики. Как это делают JKAN и другие генераторы, разобрано в главе о статических сайтах.

Платформы

Обычная веб-страница на CMS, файл README в Git-репозитории, статический сайт на GitHub Pages, общая таблица или просто листинг облачного хранилища. Когда наборов становится несколько десятков, следующий шаг — не CKAN, а статический каталог: те же файлы, но у каждого набора своя страница с описанием и разметкой.

Примеры

- awesome-public-datasets — тематический список открытых наборов с пометками о сломанных ссылках.
- Registry of Open Data on AWS — список наборов в облачном хранилище, собранный из файлов описаний.
- TLC Trip Record Data — страница выгрузки, которой пользуются годами: один издатель, понятный ряд файлов, словари полей.
- Wikimedia Downloads — дампы всех проектов Викимедиа.

Где ломается

Ссылки умирают. Это главная болезнь списков, и awesome-public-datasets с ней борется: каждой строке присвоен статус, и 12 октября 2026 года 96 записей из 1062 помечены как требующие починки. Выборочная проверка 300 случайных ссылок оттуда в тот же день: 258 открылись, 16 вернули ошибку или не ответили, ещё 26 отказали программе (коды 401, 403, 429). Последние для человека, скорее всего, живы, но робот-сборщик их уже не видит.

Нет лицензии и даты. Строка списка говорит, что данные есть, но не говорит, можно ли их использовать и когда они обновлялись.

Список дублирует чужие каталоги. Подборка ссылок на порталы и наборы, которые и так есть в каталогах, устаревает быстрее них: каталог обновляет владелец данных, список — тот, у кого нашлось время.

Список держится на одном человеке. Личный указатель на GitHub Pages или страница на сайте города перестают обновляться, как только автор уходит. У страницы «Открытые данные» на сайте ведомства при переделке сайта ломаются все адреса разом.

Как довести список до находимой публикации

1. **У каждой строки — лицензия, дата обновления, формат и размер.** Без этого список нельзя использовать, а только просматривать.
2. **Сам список — данные.** Храните его файлом CSV, JSON или YAML в репозитории и генерируйте из него страницу. Тогда список можно скачать, проверить программой и собрать в агрегатор.
3. **Ссылки проверяет робот.** Проверку ссылок по расписанию легко включить в репозиторий, например действием lychee в GitHub Actions. Сломанные строки помечайте, как в awesome-public-datasets, а не удаляйте молча.
4. **Разметка и регистрация.** Добавьте schema.org на страницу и зарегистрируйте список в каталоге своей страны или темы.

Подробно этот путь разобран в главах «Публикация без портала» и «Как сделать публикацию находимой», сводный список шагов — в чек-листах.

Когда нужен

Список — нормальный способ опубликовать десяток наборов или собрать указатель по теме. Он становится тупиком, когда нет лицензии, постоянных адресов и разметки. Заводить ради такого списка CKAN не нужно: от списка до статического каталога — один шаг, и он дешевле любого портала.

Чем отличаются от соседей

Тип	Сосед	Главное отличие
Каталог API	Портал открытых данных	описывает сервисы, а не наборы; у портала API — способ получить набор, у каталога API интерфейс и есть предмет описания
Каталог API	Каталог метаданных	реестр интерфейсов госсистем описывает, кто чем обменивается; каталог API выдаёт доступ разработчику
Маркетплейс	Портал открытых данных	на каждую выдачу — договор и условия; у портала открытых данных выдача без заявок
Маркетплейс	Пространство данных	пространство — правила и протокол обмена между участниками; маркетплейс — витрина, которая может работать поверх него
Маркетплейс	Каталог для машинного обучения	продают обучающие наборы по договору; хабы вроде Hugging Face раздают их по открытым лицензиям
Список наборов	Портал открытых данных	нет карточек, поиска и API; строка — ссылка и фраза
Список наборов	Каталог каталогов	перечисляет наборы, а не порталы; подборки awesome часто смешивают то и другое
Список наборов	Агрегатор и поисковик данных	список собирают руками, агрегатор — программой из чужих каталогов

Сводная таблица всех типов — в главе «Сравнение типов и гибриды».

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: число каталогов, владельцы, страны, платформы, доли API и ограниченного доступа, транспортные каталоги API, списки криптобирж. Запросы — в приложении.
- OpenAPI 3.2.0, 19 сентября 2025 года; AsyncAPI 3.1.0, 31 января 2026 года; RFC 9727: api-catalog, июнь 2025 года; DCAT 3 — сервис данных; API Design Rules Нидерландов
- Ответы по адресу `/.well-known/api-catalog` у одиннадцати каталогов API и HTML страницы списка API Банка Финляндии на Azure API Management — проверено 12 октября 2026 года
- Сервисы данных на data.gouv.fr: раздел API и API раздела, выгрузка всех 1265 карточек — 12 октября 2026 года; api.gouv.fr на beta.gouv.fr — перенос в data.gouv.fr и вывод из работы в 2025 году; DILA о переносе карточек, 21.11.2024; задача о переносе в репозитории data.gouv.fr

- Каталог API Норвегии и поиск сервисов данных национального каталога; таблица каталога api.gov.uk; метрики APIs.guru — проверено 12 октября 2026 года
- WSO2 API Manager; связь сервиса доступа с ресурсом в ckanext-dcat
- Официальный реестр MCP, 08.09.2025
- ODRL Information Model; Dataspace Protocol 2025-1 и выпуск 1.0.0 в Eclipse Foundation, 23.07.2025; Eclipse Dataspace Components
- Виды поставки в AWS Data Exchange и условия после окончания подписки; Databricks о переходе Delta Sharing в OpenSharing, 10.06.2026
- Rail Data Council о Rail Data Marketplace и каталог продуктов — проверено 12 октября 2026 года
- Корейские платформы больших данных: запуск десяти платформ, ZDNet Korea, 22.07.2019; шесть новых платформ, korea.kr, 22.02.2021; 16 платформ и 6842 вида данных, IT Chosun, 29.03.2022
- Биржи данных Китая: открытие Шанхайской биржи, Синьхуа, 25.11.2021; Гуйянская биржа — первая, 2015 год; обзор площадок на август 2025 года, CWW
- Объявление о переименовании OpenDataSoft в Huwise, 30 сентября 2025 года
- awesome-public-datasets и apd-core — статусы записей и выборочная проверка 300 ссылок, 12 октября 2026 года; Registry of Open Data on AWS и его файлы описаний; TLC Trip Record Data; Binance Data Collection — проверено 12 октября 2026 года
- И. Бегтин, Про рынки и биржи данных, сентябрь 2026 — четыре модели площадок и причины неудач
- И. Бегтин, Telegram: Registry of Open Data on AWS, 18.03.2021; государственные каталоги API, 04.02.2022; почему нет api.gov.ru, 02.02.2022; apistore.mos.ru, 19.11.2021; bund.dev, 24.08.2022; каталоги API как продукт для разработчиков, 10.02.2023; India Urban Data Exchange, 17.04.2024; каталоги данных компаний, 17.04.2023; открытые данные Турции, 07.02.2023
- Когда портал не нужен

Сравнение типов и гибриды

Типы каталогов различаются не названием, а главной единицей, стандартом описания и тем, без чего каталог не выполняет свою работу. Поэтому сравнивать каталог стоит с каталогами своего типа, а платформу выбирать под главный продукт.

Гибрид — сайт или организация, которые ведут каталоги нескольких типов сразу. Это частый случай: каталоги двух и более типов ведёт примерно каждое восьмое центральное ведомство и каждая восьмая международная организация. Но одной программой на все типы пользуются единицы. Практический вывод главы: назовите тип каждой части по тому, ради чего её открывают, выберите платформу под каждую часть и отдельно решите, как их связать — общей навигацией, общим поиском или сбором описаний в один каталог.

Цифры главы — по выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года), в нём 43 614 каталогов. Запросы — в приложении «Как работать с реестром». Живые сайты проверены 12 октября 2026 года.

Типы в цифрах

Тип	Каталогов	Чаще всего ведут	Свои разработки	Есть API	Ограниченный доступ	Не работают	Главные платформы
Геопортал	24 752	местные органы власти, 58%	4%	70%	1%	2,0%	ArcGIS Server, ArcGIS Hub, e-mapa, GeoServer
Научный репозиторий	8 404	университеты и научные организации, 85%	34%	63%	6%	1,4%	DSpace, Pure, IPT, Figshare
Портал открытых данных	5 967	местные органы власти, 48%	19%	69%	1%	4,1%	CKAN, OpenDataSoft (Huwise), Gosweb, Socrata
Каталог показателей	3 405	центральные органы власти, 46%	44%	37%	4%	0,9%	Swing, WordPress, Drupal, Knoema
Каталог микроданных	326	центральные органы власти, 55%	13%	70%	49%	18,1%	NADA, REDATAM, OBiBa Mica, Nesstar
Каталог метаданных	172	университеты и научные организации, 31%	38%	67%	10%	0%	FAIR Data Point, Metadata Browser, Skosmos
Поисковик данных	124	университеты и научные организации, 68%	28%	55%	1%	4,8%	OpenAIRE, VIVO
Каталог API	115	центральные органы власти, 45%	74%	51%	14%	0%	Azure API Management

Тип	Каталогов	Чаще всего ведут	Свои разработки	Есть API	Ограниченный доступ	Не работают	Главные платформы
Список наборов	107	компании, 20%	89%	14%	1%	6,5%	генераторы сайтов, CMS
ML-каталог	100	университеты и научные организации, 37%	78%	23%	5%	2,0%	генераторы сайтов, площадки соревнований
Маркетплейс	56	компании, 70%	98%	21%	91%	0%	своя разработка почти у всех

«Ограниченный доступ» — доля каталогов, у которых хотя бы часть данных выдают по регистрации, заявке или за плату. «Не работают» — каталоги, которые не открываются или официально выведены из эксплуатации. Knoema в таблице — по записям реестра: сам сервис с января 2026 года не открывается, это разобрано в главе о каталогах показателей.

Что видно из таблицы:

- **Своя разработка и API идут в разные стороны.** Где платформа общепринята — геопорталы, порталы открытых данных, микроданные, — API есть у двух третей каталогов и больше. Где каждый пишет своё — маркетплейсы, ML-каталоги, списки, — API у пятой части и меньше. Своя разработка редко доходит до программного интерфейса: его нужно не только написать, но и сопровождать.
- **Ограниченный доступ — признак типа, а не беды.** У маркетплейсов он почти всегда, у каталогов микроданных — у половины: в обследованиях сведения о людях. У портала открытых данных закрытый доступ — повод разбираться.
- **Доля мёртвых каталогов микроданных — след брошенных установок.** 47 из 59 неработающих — старые каталоги NADA статистических служб, которые ставили в рамках программ поддержки и перестали сопровождать. Разбор — в главе о каталогах микроданных.
- **Специализированная платформа почти не выходит за свой тип.** CKAN — это 1 906 порталов открытых данных и всего 39 научных репозиториев. GeoNode (770), Dataverse (257) и PxWeb (110) встречаются каждый ровно в одном типе, NADA и DSpace — с одним исключением. Девять типов из одиннадцати собирает только WordPress, семь — Drupal, по шесть — Liferay и TYPO3: универсальная CMS о типе не говорит ничего.

Чем типы отличаются по существу

Цифры показывают, как устроен рынок. Содержательная разница — в том, что каталог публикует и что он обязан сохранить.

Тип	Главная единица	Стандарт описания	Без чего не работает	Что нельзя потерять	Кому отдаёт описания
Портал открытых данных	набор: файл или таблица	DCAT	выгрузка описаний в DCAT, открытая лицензия	адрес набора и лицензию	вышестоящему порталу, общим поисковикам

Тип	Главная единица	Стандарт описания	Без чего не работает	Что нельзя потерять	Кому отдаёт описания
Геопортал	слой и картографический сервис	ISO 19115, CSW, OGC API Records	каталожный сервис и сами сервисы слоёв	возможность скачать, а не только увидеть	национальной инфраструктуре геоданных
Научный репозиторий	версия набора с DOI	DataCite	хранение файлов, DOI, версии, цитирование	версию, на которую сослались	DataCite, научным агрегаторам по OAI-PMH
Каталог показателей	показатель и временной ряд	SDMX	API рядов, методика и единицы измерения	методику ряда и его постоянный адрес	международным организациям по SDMX
Каталог микроданных	обследование и его файлы	DDI	уровни доступа, описание переменных	условия доступа и книгу кодов	архивам социальных данных
ML-каталог	набор для обучения и проверки моделей	Croissant, карточка набора	разбиение на части, лицензия, карточка	версию и разбиение, на которых меряли модель	площадкам и поисковикам по разметке
Маркетплейс	товар: набор или подписка	общего нет	договор, оплата, доставка	условия договора	никому: каталог закрыт для сбора
Поисковик или агрегатор	запись о чужом наборе	то, что собирает	сбор по протоколам, удаление дублей	ссылку на первоисточник	людям; иногда следующему агрегатору
Каталог метаданных	словарь, схема, описание каталога	SKOS, SDMX, ISO/IEC 11179	постоянные адреса понятий, версии	старые версии понятий	системам, которые на него ссылаются
Каталог API	программный интерфейс	OpenAPI	описание интерфейса, условия доступа	историю версий интерфейса	порталу открытых данных, общим каталогам API
Список наборов	ссылка на файл	чаще никакого	живые ссылки	—	обычно никому

Два вывода из таблицы. Первый: «без чего не работает» у соседних типов почти не пересекается. Портал без DCAT теряет федерацию, репозиторий без DOI — цитирование, каталог показателей без API рядов превращается в конструктор таблиц в браузере. Поэтому платформа, хорошая для одного типа, редко бывает хороша для соседнего. Второй: описания разных типов уходят в разные цепочки. Портал отдаёт DCAT национальному порталу, геопортал — ISO 19115 инфраструктуре геоданных, репозиторий — DataCite. Если сайт делает два дела, у него две цепочки, и обе нужно поддерживать. Какие возможности обязательны для какого сценария, подробно разобрано в главе «Карта возможностей».

Гибриды

Гибридом в книге называют два разных случая. Первый — **один сайт**, на котором рядом живут два продукта: наборы и слои, наборы и описания API. Второй — **одна организация**, которая ведёт несколько каталогов разных типов: статистическая служба с показателями и

микроданными, международная организация с порталом, геопорталом и библиотекой обследований. Самый развитый вариант второго случая — хаб данных: несколько каталогов под одной вывеской с общим поиском.

Гибрид — не новый тип. Каждую часть гибрида описывают и оценивают по её собственному типу: показатели статистической службы сравнивают с показателями других служб, а не с её же каталогом микроданных.

Сколько организаций ведут каталоги разных типов

В реестре 32 595 владельцев, если считать владельцем организацию с одним названием в одной стране. Каталоги двух и более типов у 2 203 из них, трёх и более — у 254. За этими владельцами 7 358 каталогов, то есть каждый шестой каталог реестра. Подсчёт по названию приблизительный: одно ведомство бывает записано по-разному, поэтому настоящих «многотипных» владельцев, скорее всего, больше.

Чем крупнее и централизованнее организация, тем чаще у неё несколько типов:

Владелец	Доля владельцев с каталогами двух и более типов
Центральные органы власти	12%
Международные организации	12%
Региональные органы власти	10%
Университеты и научные организации	7%
Местные органы власти	5%
Компании	2%

Пары типов повторяются. Среди владельцев ровно с двумя типами:

Пара типов	Владельцев	Кто обычно	Частые сочетания программ
Геопортал и портал открытых данных	953	города и регионы	ArcGIS Server и CKAN, GeoServer и CKAN, ArcGIS Hub и Socrata
Геопортал и научный репозиторий	344	университеты и научные институты	ArcGIS Server или GeoServer рядом с DSpace, THREDDS, ERDDAP
Геопортал и каталог показателей	227	статистические службы, регионы	картографический сервер рядом с базой показателей
Портал открытых данных и научный репозиторий	86	научные ведомства, университеты	CKAN рядом с Dataverse или DSpace
Каталог показателей и портал открытых данных	80	статистические службы, регионы	—
Каталог показателей и каталог микроданных	72	статистические службы	база показателей и NADA или REDATAM

Геопортал и портал открытых данных вместе ведут 1 127 организаций, если считать и тех, у кого есть третий тип. 56% из них — местные органы власти, 22% — региональные. Больше всего таких пар в США (174), Италии (90), Словакии и Испании (по 58).

Одна программа на два типа — редкость

Из 2 203 владельцев с несколькими типами всеми своими каталогами на одной программе обходятся 118, причём своя разработка здесь считается одной программой. Нагляднее всего пара «геопортал и портал открытых данных». Из 953 владельцев, у которых только эти два типа, одна и та же готовая платформа стоит на обеих сторонах у четверых: у троих ArcGIS Hub, у одного WordPress. Ещё у 39 обе части — своя разработка. Остальные держат две разные программы: картографический сервер для слоёв и портальную платформу для наборов.

Это не небрежность, а следствие таблицы из предыдущего раздела. Геопорталу нужен каталожный сервис и сервисы слоёв, порталу — DCAT и скачивание файлов. Программ, которые одинаково хорошо делают и то и другое, мало, и владельцы выбирают по специализированной программе на каждую задачу.

Пять видов гибридов

Второй продукт как дополнение

Самый частый и самый простой случай. Основной продукт один, второй его дополняет.

- **Портал открытых данных с картой** остаётся порталом, если карта показывает уже опубликованные таблицы. Он становится геопорталом, если без сервисов WMS и WFS сайту нечего показать. В реестре у 43 порталов открытых данных отмечены слои карт как содержимое.
- **Геопортал с таблицами** остаётся геопорталом, если таблицы — атрибуты слоёв или пара файлов рядом с картой.
- **Научный репозиторий с картой** остаётся репозиторием, если карта — способ найти набор по месту наблюдения, а не продукт.

Правило: тип определяет то, что получает свой адрес и свою карточку. Карта, график или предпросмотр без своего адреса — функция, а не второй каталог. Такой гибрид решается одной платформой и одним расширением: например, CKAN с ckanext-spatial для поиска по территории и ckanext-geoview для предпросмотра слоёв. Если карты становятся главным продуктом, расширения уже не хватает: в главе о CKAN это разобрано как отдельная сборка с оговоркой, когда честнее взять геоплатформу.

Одна платформа, несколько видов записей

Платформа изначально ведёт записи нескольких видов, и у каждого вида свой адрес и своя карточка. Это настоящий гибрид в одной программе.

- **data.norge.no**, национальный каталог Норвегии, держит под одной вывеской пять каталогов: наборы, API, понятия, информационные модели, государственные услуги и события. На 12 октября 2026 года в нём 9 066 наборов, 1 071 API, 6 383 понятия и 1 393 услуги. Открытыми данными отмечены только 1 644 набора из 9 066. Остальные — описания данных, которые есть у государства, но не обязательно доступны всем: по сути, это каталог метаданных о государственных данных. Работает он на собственной разработке Felles datakatalog с открытым кодом под лицензией Apache 2.0 и собирает

описания из каталогов организаций в профиле DCAT-AP-NO. По типам это портал открытых данных, каталог API и каталог метаданных сразу. Реестр учитывает его дважды — порталом и каталогом API.

- **data.gouv.fr** — 74 376 наборов, 1 265 сервисов данных (так здесь называют описания API) и 507 источников сбора. Бывший каталог API api.gouv.fr влит в портал, это разобрано в главе о каталогах API. Сервисы данных и сбор описаний из CKAN и DCAT — в ядре его программы udata (выпуск 17.12.1 от 9 октября 2026 года).
- **GeoNode** ведёт в ядре слои, карты, документы, истории на карте (GeoStories) и дашборды, а также собирает ресурсы с удалённых картографических сервисов — по документации выпуска 5.1.0 от 18 августа 2026 года. Это геопортал, внутри которого есть небольшой архив документов.
- **NADA** по описанию проекта умеет описывать не только обследования, но и показатели, геоданные, документы и изображения — в DDI Codebook, Dublin Core, ISO 19115 и 19139, IPTC. Но возможность платформы ещё не делает каталог гибридом. В библиотеке микроданных Всемирного банка все 7 145 записей — обследования, других видов нет. Каталог остаётся каталогом микроданных, а показатели и геоданные банк публикует на других платформах.
- **ArcGIS Hub** — геопортал, который ведёт себя как портал открытых данных: каждый слой можно скачать файлом, каталог отдаётся в DCAT-US и DCAT-AP. Реестр записывает Hub геопорталом 3 464 раза и порталом открытых данных — 19. Подробнее — в главе о геопорталах.

Такой гибрид работает, когда виды записей связаны между собой. В Норвегии набор ссылается на понятия, которыми описаны его поля, и на API, через который его получают. На data.gouv.fr сервис данных ссылается на набор, из которого он отдаёт строки. Если виды записей ничем не связаны, одна программа на них — просто экономия на сопровождении, и она оправдана только тогда, когда программа ведёт каждый вид не хуже специализированной.

Портал, который ещё и собирает

Портал публикует свои наборы и одновременно собирает описания из чужих каталогов. data.gouv.fr принимает публикации напрямую и забирает описания из 507 источников. data.europa.eu собирает национальные порталы, а с 1 июля 2026 года на него ведёт и адрес закрытого геопортала INSPIRE. Для пользователя разницы нет: ищет он по всему сразу. Для владельца это два разных обязательства: за свои наборы он отвечает, а за собранные — только за то, чтобы не показывать мёртвое и не размножать дубли. Чем агрегатор отличается от поисковика и как устроен сбор, разобрано в главах о поисковиках и агрегаторах и о федерации и сборе.

Несколько каталогов одного владельца

Организация ведёт каталоги разных типов на разных программах, и связывает их только навигация на сайте. Так устроено большинство статистических служб. У мексиканского INEGI в реестре 22 записи четырёх типов: банк показателей и интерактивные таблицы, таблицы на

PxWeb, служба SDMX, каталог микроданных на NADA (825 исследований на 12 октября 2026 года) и REDATAM для расчётов по переписям, картографический портал и страница открытых данных.

Это нормальная схема, если на сайте прямо сказано, что где лежит. «Показатель» и «файл обследования» — разные полки, и пользователю нужно объяснить, на какую идти. Почему для статистической службы два каталога лучше одного «универсального портала», разобрано в примере статистической службы.

Хаб

Хаб — несколько каталогов разных типов под одной вывеской, связанные общей навигацией, а часто и общим поиском. Так устроены порталы международных организаций.

- **Pacific Data Hub** Тихоокеанского сообщества по выгрузке реестра — больше 30 записей пяти типов: портал наборов на CKAN, микроданные на NADA, показатели на .Stat Suite со службой SDMX, геопорталы на GeoNode и картографическая витрина на TerriaJS, серверы THREDDS с океанографическими данными, каталог снимков STAC программы Digital Earth Pacific и демографические геопорталы PopGIS для пятнадцати стран.
- **Всемирный банк** держит отдельно показатели, библиотеку микроданных и каталог наборов, который собирает описания из остальных: на 12 октября 2026 года в нём 7 900 наборов, в том числе из библиотеки микроданных, финансовой и энергетической платформ. На главной data.worldbank.org сказано, что сайт открытых данных «расширяется до Data360», а адрес data360.worldbank.org ведёт туда же. В реестре за банком 28 записей пяти типов, большая часть — отдельные базы показателей.
- **ФАО** — каталог наборов на CKAN, больше десятка тематических карт на TerriaJS по адресам data.apps.fao.org, сайты на ArcGIS Hub и базы показателей на собственной платформе FENIX.

Хаб не делает каталоги одинаковыми. Он решает другую задачу: человеку, который пришёл за «данными о Фиджи», не нужно знать, что показатели лежат в .Stat Suite, а обследования — в NADA. Общий поиск хаба обычно строят сбором описаний из всех каталогов в один, как это делает каталог наборов Всемирного банка. Тогда у хаба появляется ещё одна цепочка сбора со всеми её проблемами.

Как определить тип гибрида

Тип каждой части гибрида определяют те же вопросы, что и тип обычного каталога, — они разобраны в главе о классификации. Для гибрида к ним добавляются четыре проверки.

1. **Что получает свой адрес.** Если у слоя, ряда или API есть своя карточка с постоянным адресом — это отдельная часть со своим типом. Если нет — это функция основного каталога.
2. **Что уходит в федерацию.** Посмотрите, что сайт отдаёт в DCAT, CSW или OAI-PMH. Если в выгрузке DCAT только наборы, а слои уходят по CSW в инфраструктуру геоданных, у сайта две цепочки и, по сути, два каталога.

3. **Какой поиск главный.** Если на главной один поиск по всему, а результаты ведут в разные программы, это хаб. Если поисков несколько и каждый ищет по своей полке, это несколько каталогов рядом.
4. **Какая программа за каждой частью.** Разные адреса на одном домене часто означают разные программы: `/catalog` — CKAN, `/microdata` — NADA, `/geoserver` — картографический сервер.

Классификацию гибрида записывают не одной строкой, а строкой на каждую часть, по шаблону из главы о классификации. Для INEGI это четыре строки: каталог показателей на своей разработке, PxWeb и службе SDMX; каталог микроданных на NADA с REDATAM; геопортал на своей разработке; портал открытых данных. Владелец, уровень и доступ у них общие, а тип, стандарт и платформа — разные.

Как устроить гибрид: три схемы

Если вам нужно два типа сразу, выбор сводится к трём схемам.

Схема	Когда подходит	Чем платите	Примеры
Одна платформа с расширениями	второй продукт маленький и его стандарт поддержан расширением или ядром	второй продукт сделан слабее, чем в специализированной программе; расширение нужно проверять при каждом обновлении	CKAN с ckanext-spatial, Dataverse для микроданных, GeoNode с документами
Две специализированные платформы и общая навигация	оба продукта важны, у каждого свой стандарт и своя цепочка	два сопровождения и два адреса; пользователю нужно объяснить, где что	статистическая служба: PxWeb или .Stat Suite и NADA; город: ArcGIS Server и CKAN
Хаб: несколько платформ и общий поиск	каталогов три и больше, и пользователю важно искать по всем сразу	ещё одна цепочка сбора, дубли, сопоставление идентификаторов, центральная команда	Pacific Data Hub, каталог наборов Всемирного банка, data.norge.no

Что специализированные платформы умеют за пределами своего типа:

Платформа	Свой тип	Второй продукт	Как
CKAN 2.12	портал открытых данных	поиск по территории и сбор описаний геоданных по CSW	расширение ckanext-spatial
CKAN 2.12	портал открытых данных	предпросмотр слоёв WMS и GeoJSON	расширение ckanext-geoview, выпуск 0.3.2 от 2 октября 2026 года
CKAN 2.12	портал открытых данных	сервисы слоёв, уровни доступа к микроданным, SDMX	нет: нужна вторая программа
ArcGIS Hub	геопортал	скачивание файлами и выгрузка в DCAT, как у портала открытых данных	ядро
GeoNode 5.1	геопортал	документы, истории на карте, дашборды	ядро
Dataverse	научный репозиторий	микроданные: описание в DDI, файлы с ограниченным доступом	ядро
Dataverse	научный репозиторий	геоданные: блок описания охвата	ядро; карта — подключаемые просмотрщики

Платформа	Свой тип	Второй продукт	Как
NADA	каталог микроданных	показатели, геоданные, документы	ядро, но без службы SDMX и сервисов слоёв
udata 17	портал открытых данных	каталог API (сервисы данных) и сбор чужих каталогов	ядро

Подробно возможности каждой платформы разобраны в карточках платформ, а сводные таблицы — в приложении.

Пример карточки решения: регион с картами и наборами

Контекст. Регион ведёт портал открытых данных на CKAN и отдаёт описания национальному порталу в DCAT-AP. Отдельно у регионального ведомства по земле работает картографический сервер со слоями и каталог GeoNetwork, который отдаёт описания по CSW в национальную инфраструктуру геоданных. Жители жалуются, что слои не находятся на портале, а руководство предлагает «свести всё в одно».

Варианты.

1. Перенести слои в CKAN файлами и закрыть геокаталог.
2. Заменить оба сайта одной геоплатформой, которая умеет и наборы, — например, ArcGIS Hub.
3. Оставить обе системы. CKAN с ckanext-spatial собирает описания слоёв из GeoNetwork по CSW и показывает их в общем поиске со ссылкой на сервисы. В национальный портал слои уходят только по одной цепочке.

Критерии. Не потерять обязательства перед инфраструктурой геоданных. Один поиск для жителя. Не породить дубли в национальном и европейском каталогах. Сколько программ сопровождать.

Решение. Вариант 3.

Почему. Вариант 1 ломает обязательство: инфраструктура геоданных ждёт описаний в ISO 19115 и сервисов слоёв, а не файлов. Вариант 2 возможен, но меняет обе системы сразу и привязывает регион к облаку одного поставщика; так делают трое из 953 владельцев этой пары в реестре. Большинство держит две программы — чаще всего картографический сервер и CKAN. Вариант 3 оставляет каждую программу на своём месте, а жителю даёт один поиск.

Последствия. На CKAN появляется ещё одно расширение, которое нужно проверять при обновлениях: ckanext-spatial на 2.12 проходит автоматические тесты, это записано в главе о CKAN. Собранные описания слоёв нужно исключить из выгрузки DCAT, если национальный портал сам собирает геокаталог, — иначе один слой дойдёт до data.europa.eu дважды. Как рвутся и путаются такие цепочки, разобрано в главе о федерации.

Когда пересмотреть. Если слои станут главным продуктом региона, а наборы — приложением к ним. Если национальный портал перестанет собирать геокаталоги напрямую. Если ckanext-spatial перестанут выпускать под новую версию CKAN.

Где ломаются гибриды

Дубли в федерации. Один и тот же слой уходит наверх дважды: через геопортал по CSW и через портал открытых данных по DCAT. Агрегатор показывает две записи с разными адресами, и обе выглядят основными. Чтобы этого не было, у каждой части гибрида должна быть своя цепочка, а собранное из соседней части не отдают наверх повторно.

Одна часть перевешивает всё. Если обе части подключены к федерации, счёт определяет та, что режет данные мельче. На data.europa.eu из 850 тысяч немецких наборов 681 тысяча приходит из одного геокаталога, почти 450 тысяч из них — планы застройки отдельных участков. Сравнивая страны или организации, смотрите, из какой части гибрида пришли записи.

Видна только одна часть. Обратная беда: в федерацию подключили одну часть гибрида и забыли другую. В июне 2024 года data.europa.eu видел из Эстонии 324 набора, потому что собирал только национальный геопортал. После подключения портала открытых данных их стало 8,7 тысячи.

Одна вещь — разные карточки. Набор в портале и слой в геопортале описывают одни данные, но с разными названиями, лицензиями и датами обновления. Пользователь не понимает, какая карточка верная, а поисковик видит две.

Два поиска без объяснения. Сайт статистической службы с поиском по показателям и отдельным поиском по обследованиям, где ни один не говорит о существовании другого. Человек ищет обследование в поиске показателей, ничего не находит и уходит.

Универсальная платформа делает второй продукт плохо. CKAN, в который положили слои файлами, не даёт сервисов WMS и WFS. Каталог показателей на CMS отдаёт таблицы картинками и PDF. Возможность «умеет и это» в описании платформы не означает, что умеет так, как нужно второму типу.

Хаб закрыт для программ. Вся ценность хаба — в связях между каталогами, но некоторые хабы закрывают их от программ. 12 октября 2026 года API портала Pacific Data Hub, его GeoNode и каталог ФАО на CKAN отвечали программам проверкой Cloudflare, а таблицы INEGI на PxWeb — ошибкой 403. Человек в браузере данные видит, а агрегатор и исследователь со скриптом — нет. Как это отражается на находимости, разобрано в главе о находимости.

Одна часть хаба умирает, остальные живут. Компоненты хаба стареют по-разному. У 110 владельцев, которые ведут и показатели, и микроданные, в реестре записаны 60 каталогов показателей на Knoema — сервисе, который с января 2026 года не открывается. Хаб нужно сопровождать как набор отдельных программ, у каждой из которых свой срок жизни.

Тип по одной записи. Агрегаторы и реестры записывают гибрид одним типом, и счёт типов от этого плывёт. ArcGIS Hub в реестре почти всегда геопортал, data.norge.no — портал открытых данных и каталог API одновременно. Сравнивая цифры по типам, помните, что граница между ними проведена человеком, а не программой.

Как проверить гибрид

Посмотреть, что на самом деле лежит внутри, можно запросами к API. Все команды проверены 12 октября 2026 года.

```
# data.gouv.fr: наборы, сервисы данных (API) и источники сбора в одной программе
curl -s https://www.data.gouv.fr/api/1/site/ | jq '.metrics | {datasets, dataservices,
harvesters}'

# data.norge.no: сколько записей в каждом из каталогов
for t in datasets data-services concepts services; do
  printf "%s: " "$t"
  curl -s -X POST -H 'Content-Type: application/json' \
    -d '{"pagination":{"size":1,"page":0}}' \
    "https://search.api.fellesdatakatalog.digdir.no/search/$t" | jq .page.totalElements
done

# NADA: какие виды записей реально есть в каталоге
for t in survey geospatial timeseries table document; do
  printf "%s: " "$t"
  curl -s "https://microdata.worldbank.org/index.php/api/catalog/search?ps=1&type=$t" | jq
.result.found
done
```

Для библиотеки микроданных Всемирного банка последняя команда даёт 7 145 обследований и ноль записей остальных видов. Старые версии NADA параметр `type` не понимают и на любой вид отвечают одним и тем же числом — так отвечает каталог INEGI. Тогда виды записей смотрят в самом каталоге, в фильтре по типу данных.

Для CKAN полезно сравнить выгрузку DCAT с поиском: если в поиске есть слои, а в `/catalog.ttl` их нет, слои собраны из геокаталога и наверх уходят по другой цепочке. Команды для отдельных типов — в разделах «Как проверить» глав о каждом типе и в чек-листах.

Правило

Назовите тип каждой части по тому, ради чего её открывают. Платформу выбирайте под эту часть, а не «максимально универсальную». Вторую задачу закройте расширением, если она маленькая, отдельной программой, если она важна, или честно урежьте. Связь между частями — навигация, общий поиск или сбор описаний — это отдельное решение со своей ценой, и его стоит записать в карточку решения, как в примере выше.

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: типы, владельцы, платформы, доли API, ограниченного доступа и неработающих каталогов, владельцы с каталогами нескольких типов и сочетания программ. Запросы — в приложении.
- data.norge.no и API поиска Felles datakatalog, код портала; data.gouv.fr, udata 17.12.1; библиотека микроданных Всемирного банка, каталог наборов Всемирного банка, data.worldbank.org; каталог микроданных INEGI; Pacific Data Hub — проверено 12 октября 2026 года.

- NADA — виды записей и стандарты описания; документация GeoNode и выпуск 5.1.0; ckanext-spatial, ckanext-geoview — проверено 12 октября 2026 года.
- И. Бегтин, Общедоступные каталоги данных международных организаций, январь 2023.
- Как классифицировать каталоги, Карта возможностей, Федерация и сбор, Разобранные примеры, Как обосновать решение.

ЧАСТЬ IV.

Возможности каталогов

Какие возможности бывают у каталогов данных, зачем они нужны и когда без них можно обойтись: экспорт метаданных, хранилище, DOI, цитирование, полнота метаданных, скорость развёртывания и другие.

1. Карта возможностей
2. Экспорт метаданных
3. Полнота и детальность метаданных
4. API и массовая выгрузка
5. Внутреннее хранилище
6. Постоянные идентификаторы, DOI и версии
7. Экспорт цитирования
8. Доступ, роли и процессы
9. Скорость развёртывания и сопровождение
10. Другие возможности

Карта возможностей

Порталы различаются не названием, а набором возможностей. Эта глава — карта всей части: какие возможности бывают, какие нужны любому каталогу, какие включает тип каталога или конкретная задача и у каких платформ каждая возможность есть в ядре, в расширении или нигде. Подробности о каждой возможности — в отдельных главах.

Коротко:

- **Шесть вещей нужны любому каталогу**, даже статическому сайту из десятка наборов: постоянный адрес, лицензия у каждого набора, дата обновления и контакт, выгрузка всех описаний, разметка для поисковиков и доступ к описаниям без регистрации.
- **Остальное включает тип каталога.** Научному репозиторию нужны DOI и версии, геопорталу — каталожный сервис, каталогу показателей — SDMX, каталогу микроданных — закрытые файлы при открытом описании, ML-каталогу — версии и описание в Croissant.
- **Тип — не единственная причина.** Вышестоящий портал, грантодатель, закон, число наборов и публикаторов делают возможность обязательной для каталога любого типа.
- **Обязательную возможность выгоднее брать из ядра платформы.** На живых каталогах она есть у большинства установок, если она в ядре, примерно у трети — если её даёт расширение, и у единиц процентов самописных сайтов.

Как пользоваться картой

1. Отметьте минимум: он одинаков для всех.
2. Найдите свой тип каталога в таблице «Что нужно в каких случаях» и выпишите строки «обязательно» и «желательно».
3. Пройдите по внешним причинам: каждая может поднять возможность с «желательно» до «обязательно».
4. Для каждой обязательной возможности посмотрите, у каких платформ она в ядре, у каких в расширении и у каких её нет.
5. Запишите и отказы, с причиной. Как превратить список в техническое задание и выбор платформы — в главах «Принципы выбора набора функций» и «От функций к платформе».

Если тип каталога ещё не ясен, начните с маршрутной карты или с классификации каталогов.

Что такое возможность

Возможность каталога — то, что каталог умеет делать для кого-то: показать набор человеку, отдать описания программе, выдать идентификатор для статьи, закрыть файл от посторонних. Это не пункт меню и не название модуля. Одна и та же возможность устроена по-разному: DOI может выдавать сама платформа, расширение к ней или чужой репозиторий, куда кладут копию.

У каждой возможности есть адресат — тот, чью задачу она решает:

- **человек на сайте** — поиск, предпросмотр, понятное описание, кнопка «Цитировать»;
- **программа** — вышестоящий портал, поисковик данных, сборщик, сервис на данных. Ей нужны выгрузки, протоколы и API;
- **владелец и редакция** — роли, проверка перед публикацией, статистика, цена сопровождения.

Каталог, сделанный только для первого адресата, хорошо выглядит, но его не видят ни агрегаторы, ни поисковики. Поэтому в минимуме ниже половина пунктов — для программ.

Группы возможностей

Группа	Что входит	На какой вопрос отвечает	Глава
Описание данных	обязательные и предметные поля, свои поля, словарь полей таблицы	можно ли по описанию понять, что внутри и можно ли этим пользоваться	Полнота метаданных
Обмен описаниями	DCAT, schema.org, отраслевые форматы: ISO 19115, DataCite, DDI, SDMX, Croissant	смогут ли другие системы прочитать описания	Экспорт метаданных
Машинный доступ	API поиска, выгрузка всех описаний, доступ к части данных, лента изменений	может ли программа забрать каталог и данные без человека	API и выгрузка
Хранение	файлы в каталоге, ссылки наружу, таблица в базе	доступен ли файл, пока жив каталог, а не сайт источника	Хранилище
Цитируемость	постоянный адрес, DOI, версии, готовая ссылка	откроет ли ссылка тот же набор через годы	Идентификаторы, Цитирование
Управление	организации, роли, модерация, закрытые файлы, регистрация	кто публикует, кто проверяет, кто видит файл	Доступ и процессы
Эксплуатация	способ запуска, обновления, расширения, люди	сколько стоит запустить и сколько — содержать годами	Развёртывание
Прочее	поиск и фильтры, предпросмотр, языки, сбор чужих каталогов, статистика, обратная связь	удобно ли людям и растёт ли каталог вместе с задачей	Другие возможности

Минимум для любого каталога

Эти возможности нужны и национальному порталу, и странице исследовательской группы. Без них каталог либо не виден, либо бесполезен.

Что	Зачем	Глава
Постоянный адрес набора и файла	на набор ссылаются статьи, отчёты и другие каталоги; адрес не меняется при правке описания и переезде сайта	Идентификаторы
Лицензия у каждого набора	без неё набор нельзя законно использовать, даже если он скачивается	Лицензии
Дата обновления и контакт	по дате видно, жив ли набор; по контакту есть кому сообщить об ошибке	Полнота метаданных
Выгрузка всех описаний	по ней каталог собирают вышестоящий портал и агрегаторы; листать поиск по страницам им неудобно и ненадёжно	API и выгрузка

Что	Зачем	Глава
Разметка schema.org и карта сайта	по ним набор находят обычные поисковики и Google Dataset Search	Экспорт метаданных
Описания без регистрации и капчи	иначе автоматический сбор не работает или ломается при каждом обновлении защиты	API и выгрузка

Выгрузка может быть простой. У небольшого сайта это один файл `data.json` или `datapackage.json`, а не полный профиль DCAT-AP. Тот же минимум записан первым разделом чек-листа требований.

Каталог без этого минимума — частая история и у крупных порталов. В ноябре 2024 года портал Москвы data.mos.ru не отдавал ни разметку schema.org, ни DCAT, и его наборов не было в поиске по данным, а небольшие общественные порталы с разметкой стояли в выдаче первыми. Подробнее — в главе об экспорте метаданных.

Что нужно в каких случаях

«Обязательно» значит: без этого каталог не выполнит свою задачу. «Желательно» — заметно помогает, но каталог полезен и без этого. «Не нужно» — для этого типа возможность лишняя или её закрывает другой инструмент. «От сотен наборов» — при десятках хватает страниц и файла описаний; свой API поиска заводят, когда наборов становится много. Те же пороги записаны в чек-листе требований.

Возможность	Портал открытых данных	Научный репозиторий	Геопортал	Каталог показателей	Каталог микроданных	ML-каталог	Небольшой сайт
Выгрузка всех описаний	обязательно	обязательно	обязательно	обязательно	обязательно	обязательно	желательно
Разметка schema.org	обязательно	обязательно	желательно	желательно	желательно	обязательно	обязательно
Экспорт в DCAT	обязательно	желательно	желательно	желательно	желательно	не нужно	желательно
Отраслевой формат	не нужно	обязательно: DataCite и OAI-PMH	обязательно: ISO 19115 и CSW или OGC API Records	обязательно: SDMX или машинный доступ к таблицам	обязательно: DDI	обязательно: Croissant	не нужно
API поиска	от сотен наборов	от сотен наборов	от сотен наборов	желательно	от сотен наборов	от сотен наборов	не нужно
Доступ к части данных без скачивания	желательно	не нужно	обязательно	обязательно	желательно	желательно	не нужно
Файлы в самом каталоге	желательно	обязательно	не нужно	не нужно	желательно	желательно	не нужно
DOI	не нужно	обязательно	не нужно	не нужно	желательно	желательно	не нужно
Версии	желательно	обязательно	не нужно	желательно	желательно	обязательно	не нужно
Экспорт цитирования	желательно	обязательно	не нужно	желательно	желательно	желательно	желательно
Закрытые файлы при открытом описании	не нужно	желательно	не нужно	не нужно	обязательно	желательно	не нужно

Почему так:

- **Портал открытых данных** живёт в цепочке «город → регион → страна → общий поисковик». Без DCAT и выгрузки описаний его не соберёт вышестоящий портал. API поиска нужен, когда наборов сотни и больше: городу с тридцатью наборами хватает страниц и файла описаний. Файлы у себя желательны, потому что внешние ссылки умирают: при полной выгрузке data.gov.uk в 2023 году около 10% ссылок на файлы не открывались. DOI здесь обычно избыточен: данные обновляются, и вместо версий с идентификаторами достаточно архива прошлых файлов.
- **Научный репозиторий** нужен, чтобы на данные сослались в статье. Поэтому обязательны DOI, версии и готовая строка цитирования — и у маленького репозитория тоже, если на версию ссылаются. Файлы должны лежать в самом репозитории: ссылка в статье не может зависеть от чужого сайта. Закрытые файлы нужны не всем, но часто: эмбарго до выхода статьи, данные о пациентах.
- **Геопортал** описывает слои и сервисы. Отраслевой язык — ISO 19115 и каталожный сервис CSW или OGC API Records; это обязательно и при небольшом числе слоёв. Доступ к части данных — это картографические сервисы WMS, WFS и OGC API: без них слой можно только посмотреть на карте. Хранить файлы у себя не обязательно: данные обычно уже лежат в картографическом сервере.
- **Каталог показателей** отдаёт не файлы, а показатели с разрезами, поэтому главное — выборка по годам и территориям. Без SDMX или хотя бы машинного доступа к таблицам с описанием показателя каталог остаётся конструктором в браузере. Выгрузку описаний здесь заменяет полный список таблиц или рядов с методикой. Версии желательны: показатели пересматривают, и аналитику нужно знать, на какой выпуск он опирался.
- **Каталог микроданных** описывает обследования, а сами записи о людях выдаёт по уровням доступа: описание в DDI видно всем, файл — по заявке. Хранить файлы у себя нужно, если вы сами их выдаёте; библиотека микроданных Всемирного банка в основном ведёт во внешние репозитории. Доступ к части данных здесь — онлайн-таблицы, которые считают распределения, не выдавая записей, как в REDATAM.
- **ML-каталог** отдаёт наборы, на которых сравнивают модели. Результаты сравнимы, только если набор и его деление на обучающую и тестовую части не менялись, поэтому версии обязательны. Описание в Croissant требует трек данных конференции NeurIPS у каждой подачи с новым набором, а Kaggle, OpenML, Hugging Face и Dataverse делают его сами. Закрытые файлы — это доступ по запросу, как у части наборов на Hugging Face.
- **Небольшой сайт** с оперативными данными обходится без API поиска и без DOI. Поисковикам нужна разметка schema.org, иначе набор не найдут. Если на данные нужно сослаться как на неизменную версию, это уже задача научного репозитория: DOI обязателен при любом размере, и проще всего получить его в общем репозитории.

Что делает возможность обязательной для любого типа

Таблица выше — ориентир, а не закон. Региональный портал, который обязан отдавать метаданные национальному, делает DCAT обязательным, даже если для его типа стоит «желательно». Ниже — причины, которые поднимают возможность до «обязательно» независимо от типа каталога.

Если	То обязательно	Глава
Ваши описания собирает вышестоящий портал	DCAT в профиле, который принят в стране (DCAT-AP, DCAT-US), и устойчивый адрес выгрузки	Экспорт метаданных
Данные входят в наборы высокой ценности ЕС (регламент 2023/138)	API и массовая выгрузка, лицензия не строже CC BY 4.0	API и выгрузка
Слои входят в инфраструктуру пространственных данных страны	описания по ISO 19115 и каталожный сервис	Геопорталы
Грантодатель или журнал требует ссылку на данные	DOI, версии, экспорт цитирования	Идентификаторы
Набор сопровождает статью в треке данных NeurIPS	описание в Croissant	ML-каталоги
Наборов сотни и больше	API поиска, поиск по тексту описаний и фильтры	Другие возможности
Публикует больше одной команды	организации и роли	Доступ и процессы
Публикуют внешние люди	модерация до выхода в поиск, регистрация по приглашению	Доступ и процессы
Данные персональные, лицензируемые или чувствительные	закрытые файлы при открытом описании, журнал заявок	Доступ и процессы
В стране несколько официальных языков или аудитория международная	языки интерфейса и описаний	Другие возможности
Вы собираете чужие каталоги	сбор по расписанию из DCAT, API CKAN, CSW, OAI-PMH	Федерации и агрегация
Закон или отраслевой справочник требует реквизит, которого нет в схеме	свои поля описания	Полнота метаданных
Каталог должен прожить десять лет	файлы у себя, платформа с понятной политикой обновлений, план закрытия	Развёртывание

Работает и обратное: возможность из столбца «обязательно» можно вычеркнуть, если её закрывает кто-то другой. Геопортал не держит у себя файлы, если слои живут в картографическом сервере. Исследовательская группа не заводит свои DOI, если кладёт выпуск в Zenodo. Город не делает свой API поиска, если его наборы уже собирает региональный портал.

Где возможность живёт в платформах

Одна и та же возможность в одной платформе есть в ядре, в другой появляется только с расширением, в третьей её нет совсем. В таблицах этой части четыре пометки:

- **ядро** — возможность есть в самой платформе;
- **расширение** — её добавляет отдельный модуль, у которого свой автор и свои риски при обновлении;
- **нет** — платформа этого не умеет, нужна другая платформа или своя разработка;
- **не проверено** — по документации октября 2026 года клетку нельзя было закрыть.

Сводка по главам этой части:

Возможность	В ядре	Через расширение или настройку	Нет
Экспорт в DCAT	DKAN, udata, Piveau, Huwise (OpenDataSoft), Socrata, ArcGIS Hub, InvenioRDM, GeoNetwork	CKAN — ckanext-dcat; у Dataverse — сторонний экспортёр DCAT 3	DSpace, NADA, PxWeb, GeoNode; у Dataverse в ядре экспорта DCAT нет
Разметка schema.org	Dataverse, InvenioRDM, GeoNetwork, NADA, Socrata; у Huwise в ядре, но включает администратор	CKAN — ckanext-dcat	PxWeb; у udata, Piveau, ArcGIS Hub и GeoNode в сверке 13 октября 2026 года разметки на странице нет; на статическом сайте блок пишет автор шаблона
Выгрузка всех описаний	Dataverse, InvenioRDM, DSpace (OAI-PMH); GeoNetwork (CSW); DKAN и облачные сервисы (DCAT); JKAN (файл `data.json`)	CKAN — ckanext-dcat	статический сайт, пока автор не собрал файл описаний
API поиска	CKAN, DKAN, облачные сервисы, Dataverse, InvenioRDM, DSpace, GeoNetwork, NADA, PxWeb	—	JKAN, статический сайт
Файлы в каталоге	CKAN, Dataverse, InvenioRDM, DSpace, NADA, облачные сервисы	у CKAN облачное хранилище вместо диска — настройкой или расширением	PxWeb и .Stat Suite: хранят таблицы, а не файлы
Доступ к части данных	CKAN (DataStore), облачные сервисы, ArcGIS Hub, Dataverse (табличные файлы), PxWeb, .Stat Suite	у CKAN заливку CSV в DataStore делает расширение XLoader	GeoNetwork (ссылается на чужие сервисы), статический сайт
DOI	Dataverse, InvenioRDM, Figshare; DSpace и GeoNetwork — после включения в настройках. Везде нужен договор с DataCite	CKAN — ckanext-doi; документация называет только CKAN 2.9 и 2.10, на 2.11 и 2.12 не тестировалось	DKAN, Huwise (OpenDataSoft), Socrata, ArcGIS Hub, PxWeb; у NADA только поле для DOI, выпущенного отдельно
Версии	Dataverse, InvenioRDM, Figshare, DSpace	—	CKAN (журнал изменений — не версия), облачные порталы открытых данных
Экспорт цитирования	Dataverse, InvenioRDM, Figshare, GeoNetwork 4.4; NADA — через DataCite, если у исследования есть DOI	CKAN — ckanext-doi (только при DOI) или молодое ckanext-citeproc	DSpace до 10.1 включительно (только метатеги для Google Scholar), облачные порталы открытых данных
Свои поля описания	Dataverse, InvenioRDM, DSpace, DKAN (замена схемы); Huwise (OpenDataSoft) и Socrata — шаблоны полей в настройках	CKAN — ckanext-scheming; GeoNetwork — плагины схем для профилей стран	ArcGIS Hub: один стиль из готовых стандартов на организацию; JKAN и статический сайт
Закрытые файлы при открытом описании	Dataverse, InvenioRDM, DSpace, NADA, Huwise (OpenDataSoft)	у CKAN в ядре закрывается только набор целиком, вместе с описанием; тонкая настройка — расширения	статический сайт; у Socrata закрытый набор скрыт целиком
Сбор чужих каталогов	Piveau; GeoNetwork (CSW, OAI-PMH, сервисы OGC, коллекции STAC и другие); Dataverse (клиент OAI-PMH); udata (шесть сборщиков в поставке, по умолчанию выключены); Huwise (сборщики CKAN, Socrata, ArcGIS, data.json и других; CSW включают по запросу); Socrata (Catalog Connector: data.json и публичный ArcGIS Server)	CKAN — ckanext-harvest, ckanext-dcat, ckanext-spatial	JKAN, статический сайт; у ArcGIS Hub в сверке октября 2026 года не проверено

Кто ведёт каждое расширение CKAN и с какими версиями ядра оно работает, записано в главе «CKAN: ядро и расширения». Полная сводка по платформам — в приложении, карточки платформ — в главе «Карточки платформ».

Пометка «ядро» не значит «включено». DOI в Dataverse, DSpace и GeoNetwork заработает только после договора с DataCite и настройки, а предпросмотр в Dataverse — после того, как администратор поставит просмотрщики. Проверяйте конкретный сайт, а не только документацию платформы.

Как это выглядит на живых каталогах

Разница между ядром и расширением видна не только в документации. Реестр dataportals-registry записывает, какие интерфейсы удалось найти у каждого каталога. Ниже — сколько каталогов отдают возможность, обязательную для их типа, в зависимости от платформы. Выпуск реестра 1.27.0, все записи типа, включая недействующие; научные репозитории посчитаны вместе с универсальными.

Возможность и тип каталога	Платформы, где она в ядре	Где её даёт расширение или настройка	Свои разработки
DCAT у порталов открытых данных	Socrata — 239 из 245, OpenDataSoft — 348 из 454, DKAN — 103 из 161	CKAN с ckanext-dcat — 633 из 1906	24 из 1144
DOI у научных репозиториев	Figshare — 266 из 288, Dataverse — 172 из 257	DSpace, где DOI включают настройкой, — 218 из 829	158 из 2840
OAI-PMH у научных репозиториев	DSpace — 665 из 829, Dataverse — 177 из 257	—	19 из 2840
CSW или OGC API Records у геопорталов	ArcGIS Hub — 3257 из 3464, GeoNetwork — 686 из 762, GeoNode — 628 из 770	—	4 из 1043

Что из этого следует:

- Если возможность в ядре, она есть примерно у двух третей установок и больше, а у облачных Socrata и ArcGIS Hub — почти у всех: там её не нужно ни ставить, ни включать.
- Если её нужно поставить отдельно или включить настройкой, она есть у трети или четверти: ckanext-dcat у CKAN, DOI у DSpace. У Dataverse DOI в ядре, но платформа умеет выдавать и другие идентификаторы — Handle и постоянные ссылки самой установки, поэтому DOI отмечен у двух третей сайтов.
- У своих разработок обязательной возможности почти нет: DCAT отдают 2% самописных порталов открытых данных, OAI-PMH — меньше 1% самописных репозиториев.
- Картографический сервер — ещё не каталог. Из 4674 геопорталов на ArcGIS Server каталожный сервис найден у одного, из 1927 муниципальных порталов e-mapa — ни у одного: они показывают слои, но описаний для сбора не отдают.

По всем каталогам картина скромнее, чем требует таблица «Что нужно в каких случаях». DCAT найден у 26% порталов открытых данных, OAI-PMH — у 25% научных репозиториев, DOI — у 17%, каталожный сервис — у 20% геопорталов, SDMX — у 5,2% каталогов показателей. Из этих 177 каталогов показателей 113 — витрины Knoema, а сам Knoema с января 2026 года

не открывается (подробнее в главе о каталогах показателей). То, что карта называет обязательным, на деле есть у меньшинства, и обычно причина в платформе: возможность не дали ни ядро, ни поставленное расширение.

Числа — нижняя граница: реестр отмечает интерфейс, только если нашёл его по известным адресам. Разметку schema.org реестр почти не проверяет, поэтому её в таблице нет. Запросы — в приложении «Как работать с реестром».

Как закрыть возможности без портала

Многие возможности из карты не требуют портала. Статический сайт, Git-репозиторий, общий научный репозиторий вроде Zenodo и облачный хаб закрывают весь минимум и часть остального.

Возможность	Как закрыть без портала	Чего не будет
Постоянный адрес	свой домен с неизменной структурой адресов; DOI выпуска в Zenodo	—
Лицензия	файл `LICENSE` и поле лицензии в `datapackage.json`	—
Выгрузка всех описаний	файл `data.json` или `datapackage.json` рядом со страницами; генератор вроде JKAN собирает его сам	протоколов сбора вроде OAI-PMH и CSW
Разметка schema.org	блок JSON-LD в шаблоне страницы набора	проверки полей: за ней следит автор шаблона
API поиска	при десятках наборов не нужен	поиска для программ
Доступ к части данных	Datasette — программа, которая одной командой публикует базу SQLite с веб-страницей и JSON API	чистой статики: Datasette работает как небольшой сервер
Файлы	рядом со страницами, во вложениях к выпуску, в облачном хранилище	больших файлов в Git: GitHub не принимает файлы больше 100 МиБ
DOI и версии	метки выпусков в репозитории и автоматическая архивация выпусков в Zenodo	—
Экспорт цитирования	файл `CITATION.cff`: GitHub показывает по нему кнопку для ссылки на репозиторий	—
Роли и проверка	права на репозиторий и проверка изменений перед слиянием	модерации внешних публикаторов
Закрытые файлы	—	так публикуют только открытое
Сбор чужих каталогов	—	его не будет

Если таблица «Что нужно» для вашего случая требует только минимума, портал скорее всего не нужен. Когда так бывает и как сделать такую публикацию находимой — в части «Публикация без портала», особенно в главах о Git-репозиториях и о находимости.

Как быстро проверить живой портал

Несколько запросов за десять минут показывают, какие возможности из карты есть у конкретного сайта. Подставьте адрес портала и адрес страницы любого набора:

```
P=https://portal.example
```

```
# Выгрузка всех описаний: DCAT-US и DCAT в RDF
curl -sI "$P/data.json" | head -1
curl -sI "$P/catalog.ttl" | head -1

# Карта сайта
curl -sI "$P/sitemap.xml" | head -1

# Разметка schema.org на странице набора: число блоков JSON-LD
curl -s "$P/dataset/some-dataset" | grep -c 'application/ld+json'

# API поиска CKAN и список установленных расширений
curl -s "$P/api/3/action/package_search?rows=0"
curl -s "$P/api/3/action/status_show"

# Научный репозиторий и геопортал
curl -sI "$P/oai?verb=Identify" | head -1
curl -sI "$P/csw?service=CSW&request=GetCapabilities" | head -1
```

Ответ `200` — хороший знак, `404` по всем адресам выгрузки — повод искать её в документации портала или считать, что её нет. Дальше три проверки глазами: есть ли у пяти случайных наборов лицензия и дата обновления; открываются ли файлы; виден ли DOI и блок «Как цитировать», если портал их обещает. Подробные проверки — в разделе «Как проверить живой портал» каждой главы этой части и в сводных чек-листах.

Типичные ошибки

- **«Обязательно» на всё.** Если обязательно всё, сравнение платформ теряет смысл, а цена проекта растёт за счёт возможностей, которыми никто не пользуется.
- **Список из чужого задания.** Национальному порталу нужны сбор чужих каталогов, роли для сотни ведомств и витрина примеров. Городу с тридцатью наборами — нет.
- **Витрина вместо минимума.** Поиск с фильтрами и графики на главной делают до выгрузки описаний и разметки. В итоге сайтом пользуются люди, а программы его не видят.
- **Предпросмотр вместо API.** Таблица на странице помогает человеку, но её нельзя процитировать и нельзя надёжно забрать программой.
- **DOI для оперативных данных.** Расписанию транспорта, которое меняется каждый день, нужен постоянный адрес и дата обновления, а не новый DOI на каждую выгрузку.
- **Расширение без проверки.** «Есть расширение» не значит «работает на вашей версии». Перед выбором проверяют, кто его ведёт и на каких версиях ядра идут тесты, — методика в главе «Экосистемы расширений».

Как устроены главы этой части

Каждая глава о возможности устроена одинаково: что это, зачем, какие бывают варианты, когда нужно и когда нет, как это сделано в платформах, как проверить живой портал. В таблицах платформ те же четыре пометки, что выше: ядро, расширение, нет, не проверено.

Источники

- Главы этой части и документация платформ, на которую они ссылаются.

- Ориентиры по типам сверены с тем, как устроены реальные каталоги в реестре dataportals-registry. Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: доли интерфейсов и DOI по типам и платформам — запросы 15 октября 2026 года к этой выгрузке, приведены в приложении.
- DCAT 3, рекомендация W3C, schema.org/Dataset
- Регламент (ЕС) 2023/138 о наборах данных высокой ценности — API, массовая выгрузка и лицензия
- NeurIPS 2026: правила трека Evaluations & Datasets — Croissant у каждой подачи в этот трек, которая вводит новый набор
- Сбор чужих каталогов в GeoNetwork 4.4 — CSW, OAI-PMH, сервисы OGC, коллекции STAC, проверено 12 октября 2026 года
- И. Бегтин, Почему данные по России так непросто найти, ноябрь 2024 — data.mos.ru и поиск по данным; Хорошие и плохие практики публикации данных, ноябрь 2024 — доля нерабочих ссылок на data.gov.uk

Экспорт метаданных

Каталог полезен не только людям на его сайте. Другие порталы, поисковики и программы должны уметь прочитать описания наборов. Для этого каталог отдаёт метаданные в формате, который эти системы уже понимают.

Главный вывод главы: ответить по адресу `/data.json` — ещё не значит отдать описания. У выгрузки есть профиль, и профили не взаимозаменяемы: data.europa.eu ждёт DCAT-AP, data.gov — DCAT-US, Google — разметку на странице набора. Платформа обычно умеет несколько форматов сразу, но нужный профиль включают настройкой, а поля описания должны быть заполнены так, как требует именно он. Сами стандарты разобраны в главе «Стандарты метаданных». Здесь — о возможности платформы их отдавать.

Что это

Экспорт метаданных — это готовое описание набора или всего каталога в формате, который читает чужая система. У него три адресата, и под каждого свой формат.

Другие каталоги. Им нужен общий словарь описаний. Для открытых данных это DCAT — рекомендация W3C, словарь каталогов данных. По нему национальный портал забирает региональные, а общий европейский портал — национальные. Сам DCAT задаёт только смысл полей. Что из них обязательно и какими словами их называть, решает профиль — уточнение словаря под страну или отрасль:

- **DCAT-AP** — профиль для Европы. Действующая версия — 3.0.1, рекомендация SEMIC от 27 октября 2025 года. Черновик 3.0.2 от 17 июня 2026 года рекомендацией ещё не стал. Поверх него есть отраслевые уточнения: GeoDCAT-AP для геоданных, StatDCAT-AP для статистики, DCAT-AP HVD для наборов высокой ценности, HealthDCAT-AP для медицинских данных;
- **DCAT-US** — профиль для федеральных ведомств США. Ведомство публикует один файл `data.json`, а data.gov его собирает. Новая версия — DCAT-US 3.0, она построена на DCAT 3. Старую версию 1.1 data.gov ещё принимает на переходный период.

Поисковики. Им нужна разметка прямо в HTML-странице набора: блок `application/ld+json` со словарём schema.org и типом Dataset. Её читает Google Dataset Search. Это не замена DCAT: у разметки другая аудитория, и Google прямо пишет, что понимает и schema.org, и DCAT, но разметку ищет на странице набора.

Отраслевые системы. У них свои форматы и протоколы:

- геоданные — ISO 19115 и каталожный сервис CSW, а также OGC API Records;
- наука — схема DataCite и протокол OAI-PMH;
- микроданные — DDI;
- статистика — SDMX;
- машинное обучение — Croissant поверх schema.org.

Протокол и формат — разные вещи, и их часто смешивают. CSW и OAI-PMH отвечают на вопрос «как спросить каталог». DCAT, ISO 19115 и DataCite отвечают на вопрос «на каком языке придёт ответ». Один и тот же CSW умеет ответить и карточкой ISO 19115, и описанием DCAT — если это умеет программа каталога.

Зачем нужна

Без стандартной выгрузки каталог виден только тем, кто открыл его сайт. Вот кто остаётся за бортом.

Вышестоящий портал. Он собирает описания по профилю, который принят в стране, и по известным адресам. Каталог с «почти DCAT» или со своим JSON в сбор не попадёт, даже если набор полей похож. Как устроены такие цепочки и где они рвутся — в главе о федерациях и сборе.

Поисковик по данным. В ноябре 2024 года Google Dataset Search по русским запросам выдавал на первых местах наборы с hubofdata.ru и ngodata.ru — двух небольших общественных порталов на CKAN. Не потому, что они самые полные, а потому, что на их страницах есть разметка schema.org. Портал открытых данных Москвы data.mos.ru, крупный и рабочий, не отдавал ни schema.org, ни DCAT, и его наборов не было ни в Google Dataset Search, ни в поисковике Dateno. Так же в 2022 году выпадал портал Индии data.gov.in: обычный поиск Google его индексировал, поиск по данным — нет. Разметка важнее размера каталога.

Агрегаторы и архивы. Им нужна полная выгрузка, которую можно повторить: один файл или протокол с выборкой по дате. Карта сайта для этого не годится — в ней только адреса страниц, без лицензии и файлов. Чем выгрузка описаний отличается от API поиска, разобрано в главе об API и массовой выгрузке.

Закон. Для органов власти в ЕС профиль DCAT-AP — практическое требование обмена с data.europa.eu. Для наборов высокой ценности по регламенту 2023/138 к нему добавляется профиль DCAT-AP HVD.

Выгрузка не исправляет пустые поля. Если в описании нет лицензии и издателя, в DCAT их тоже не будет: программа переносит то, что заполнено. Какие поля требуют профили и как проверить, что они заполнены по существу, разобрано в главе о полноте метаданных.

Варианты реализации

Описания отдают пятью способами. Они не исключают друг друга: взрослому каталогу нужны сразу несколько.

Способ	Что получает потребитель	Сильная сторона	Слабое место
Разметка на странице	schema.org в HTML каждого набора	её читает поисковик, обходя сайт	только один набор за раз; страницу должен отдать сам сервер, а не дорисовать скрипт
Один файл на весь каталог	`data.json`, DCAT, таблица CSV	забирается одной командой	растёт вместе с каталогом; у больших каталогов становится тяжёлым

Способ	Что получает потребитель	Сильная сторона	Слабое место
Выгрузка по страницам	тот же DCAT, со ссылкой на следующую страницу	работает на каталоге любого размера	надо пройти все страницы; если каталог меняется во время обхода, записи могут потеряться
Протокол	OAI-PMH, CSW, OGC API Records	есть выборка по дате изменения	свой запрос под каждый протокол
Запись во внешний реестр	метаданные уходят в DataCite вместе с DOI	набор находится по идентификатору, без обхода сайта	только для наборов с DOI

Два решения внутри этих способов стоят отдельного разговора.

Профиль включают, а не получают «заодно». У ckanext-dcat по умолчанию включён профиль DCAT-AP 3, а DCAT-US 3, HealthDCAT-AP и schema.org — отдельные профили, которые администратор называет в настройке. Поэтому портал на CKAN с расширением может отдавать DCAT и при этом не проходить валидатор профиля своей страны. То же у облачных сервисов: сайт на ArcGIS Hub по умолчанию отдаёт DCAT-US 1.1 и DCAT-AP 2.1.1, а ленты DCAT-US 3.0 и DCAT-AP 3.0.0 включаются отдельно.

Выгрузка и страница — разные представления. Одно и то же описание можно попросить в разных форматах файла: Turtle, RDF/XML, JSON-LD. Это называется согласованием содержимого: программа указывает в заголовке `Accept`, что ей нужно, или берёт адрес с нужным окончанием. Для человека это незаметно, для сборщика важно. Но страница набора и файл выгрузки живут отдельно. 13 октября 2026 года страница набора на data.gouv.fr не содержала разметки schema.org вовсе: в HTML была только ссылка на файл Turtle. Каталог отдаёт DCAT, а поисковик по данным его страницы не прочитает.

Ещё одна ловушка — страница, которую рисует скрипт. Если описание и разметка догружаются после открытия страницы, робот, который не выполняет скрипты, видит пустую оболочку. Так устроена страница набора на data.europa.eu. Подробнее — в главе о находимости.

Когда нужна, а когда нет

Ситуация	Что нужно	Чем обычно закрывают
Портал, который собирает национальный или data.europa.eu	DCAT в профиле страны, без регистрации, с устойчивым адресом	ядро DKAN, udata, Huwise, Socrata, ArcGIS Hub; у CKAN — ckanext-dcat с нужным профилем
Наборы высокой ценности в ЕС	профиль DCAT-AP HVD поверх DCAT-AP	профиль в ckanext-dcat, форматтеры GeoNetwork, настройка ленты у Hub
Каталог, который должны находить поиском	разметка schema.org на странице каждого набора и карта сайта	ядро Dataverse, InvenioRDM, GeoNetwork, NADA; у CKAN — тот же ckanext-dcat
Геопортал в инфраструктуре страны	ISO 19115 и CSW или OGC API Records; для портала открытых данных — ещё GeoDCAT-AP	ядро GeoNetwork
Научный репозиторий	DataCite и OAI-PMH; schema.org — для поиска	ядро Dataverse, InvenioRDM, DSpace

Ситуация	Что нужно	Чем обычно закрывают
Каталог микроданных	DDI на каждое обследование; schema.org — для поиска	ядро NADA
Каталог показателей	SDMX для международных партнёров	ядро .Stat Suite; у PxWeb полноценного SDMX нет
Каталог для машинного обучения	Croissant	ядро Hugging Face, Kaggle, OpenML; у CKAN — плагин в ckanext-dcat, у Dataverse — в ядре с версии 6.10
Десяток наборов, который никто не собирает	schema.org и один файл `data.json`	статический сайт, JKAN
Служебная витрина для своих сотрудников	ничего	экспорт не нужен

Разметку schema.org стоит ставить почти всем: она ничего не стоит на типовой платформе и решает, попадёт ли набор в поиск по данным. Полноценный профиль DCAT-AP нужен только тем, кого реально собирают. Для каталога из десятка наборов он избыточен — но файл со всеми описаниями полезен и ему: по нему каталог подберёт агрегатор, когда до него дойдёт очередь.

Коротко, как выглядит обоснование по шаблону карточки решения:

- **Контекст.** Региональный портал открытых данных на CKAN 2.11. Национальный портал собирает регионы по DCAT-AP 3, а сотрудники просят «чтобы нас было видно в Google».
- **Варианты.** Поставить ckanext-dcat с профилем по умолчанию; включить в нём профиль DCAT-AP 3 и разметку schema.org; написать свою выгрузку в JSON «по мотивам DCAT».
- **Критерии.** Какой профиль ждёт национальный портал, попадает ли набор в поиск, сколько стоит сопровождение при обновлении CKAN.
- **Решение.** ckanext-dcat с профилем DCAT-AP 3 и плагином разметки. Свою выгрузку не писать.
- **Почему.** Расширение ведёт команда CKAN, оно проверяется на CKAN 2.10, 2.11 и 2.12. Свой JSON придётся объяснять каждому сборщику отдельно, а профиль по умолчанию может не совпасть с тем, что ждёт национальный портал.
- **Когда пересмотреть.** Если страна примет национальный профиль поверх DCAT-AP или если портал переедет с CKAN.

Как это сделано в платформах

Пометки — ядро, расширение, нет, не проверено, как в сводной таблице. Сверено с документацией и живыми установками 13 октября 2026 года.

Платформа	DCAT	schema.org	Отраслевой формат	Что важно знать
CKAN 2.10–2.12	расширение ckanext-dcat: DCAT-AP 2 и 3, DCAT-US 3, HealthDCAT-AP	расширение: плагин structured data в том же ckanext-dcat	расширение ckanext-dcat отдаёт Croissant; CSW — расширение ckanext-spatial, только для собранных записей	по умолчанию профиль DCAT-AP 3; выгрузка постраничная, с выборкой по дате
DKAN	ядро: файл `data.json`	не проверено	нет	схема ядра — DCAT-US 1.1; живые установки на новом DKAN отдают и DCAT-US 3.0

Платформа	DCAT	schema.org	Отраслевой формат	Что важно знать
udata	ядро: DCAT по страницам и файлом у набора	нет: в HTML страницы разметки нет, только ссылка на файл	нет	ночная выгрузка всего каталога в CSV — отдельная задача ядра
Huwise (OpenDataSoft)	ядро: каталог одним файлом в RDF, Turtle и DCAT	ядро, но включает администратор: на opendata.paris.fr разметка есть, на data.economie.gouv.fr её нет	нет	у части установок есть национальный профиль: на data.bs.ch это DCAT-AP-CH
Socrata	ядро: `data.json` одним файлом	ядро: JSON-LD на странице набора	нет	разметка на data.cityofnewyork.us содержит лицензию, издателя и список файлов
ArcGIS Hub	ядро: DCAT-US 1.1 и DCAT-AP 2.1.1; DCAT-US 3.0 и DCAT-AP 3.0.0 включаются отдельно	нет: на странице набора opendata.dc.gov разметки нет	ядро: OGC API Records	в ленту попадает всё содержимое сайта, не только наборы
Piveau	ядро: DCAT в нескольких форматах, SPARQL	нет: страница набора на data.europa.eu рисуется скриптом, в HTML разметки нет	ядро: SPARQL	на data.europa.eu описание набора в Turtle открывается по адресу API
Dataverse 6.x	нет в ядре; DCAT 3 — сторонний экспортёр для Dataverse 6	ядро: с версии 6.10 в страницу встроена краткая разметка Croissant, прежняя schema.org осталась в выгрузке	ядро: OAI-PMH, DataCite, DDI, Dublin Core	Croissant встроен в ядро в версии 6.10, раньше он был отдельным экспортёром
InvenioRDM v14, Zenodo	ядро: запись в DCAT по заголовку `Accept`	ядро: JSON-LD на странице записи	ядро: OAI-PMH, DataCite, Dublin Core, MARCXML	одного файла со всеми описаниями нет, только протокол
DSpace 7 и новее	нет	нет	ядро: OAI-PMH и Dublin Core	страница набора DSpace@MIT отвечает программам проверкой, поэтому разметку там проверить не удалось
GeoNetwork 4.4	ядро: DCAT, DCAT-AP и GeoDCAT-AP через CSW и отдельные форматтеры	ядро: JSON-LD schema.org на странице записи; поля Croissant — в ветке будущего 4.4.13	ядро: ISO 19115 и CSW; OGC API Records — отдельный сервис	у каталога EEA среди форматов CSW есть DCAT, DCAT-AP HVD и mobilityDCAT-AP
GeoNode 5.1	нет: каталог отдаёт API и CSW, но не DCAT	нет: страница каталога — оболочка, разметки в HTML нет	ядро: CSW и ISO 19115, внутри работает pycsw; карта сайта в ядре	файл `data.json` у части старых установок есть, но это не выгрузка ядра GeoNode 5
NADA	нет	ядро: JSON-LD на странице исследования	ядро: DDI Codebook по каждому исследованию; OAI-PMH нет	в разметке библиотеки микроданных Всемирного банка нет списка файлов и лицензии
PxWeb	нет	нет	ядро: описание таблиц в JSON-stat; SDMX-ML 2.0 — только в старом API	в PxWebApi 2 формата SDMX нет
JKAN	ядро: файл `data.json`, объявляет схему DCAT-US 1.1	нет: шаблон страницы разметку schema.org не ставит; на странице есть RDFa `dcat:Dataset`	нет	шаблон выводит фиксированный список полей; 14 октября 2026 года opendata.scot отдавал шесть, издателя среди них не было
Статический сайт своими руками	нет: файл кладёт автор	нет: блок в шаблоне пишет автор	нет	готовый шаблон разметки есть в книге

Три строки в таблице расходятся с тем, что обычно говорят о платформах.

GeoNode не отдаёт DCAT. Его каталог — это pycsw: CSW отвечает карточками ISO 19115, Dublin Core и ещё несколькими форматами, среди которых DCAT нет. Так отвечали живые установки GeoNode в Никарагуа и Боливии 13 октября 2026 года. Файл `data.json`, который реестр находит у части установок GeoNode, к ядру версии 5.1 не относится: в её коде такого адреса нет, а открывшийся файл содержал свой список полей, а не каталог DCAT-US.

udata не размечает страницы. data.gouv.fr на udata 17.12.1 отдаёт описание набора файлом Turtle, но в HTML страницы блока schema.org нет. Для сборщика каталог открыт, для Google Dataset Search — нет.

У Dataverse разметка сменилась. До версии 6.10 в страницу встраивалась schema.org, и её можно вернуть настройкой. С версии 6.10 на её месте краткая разметка Croissant: без списка файлов, зато её просит Google, чтобы набор попал в фильтр «с поддержкой Croissant». Полную разметку со списком файлов по-прежнему отдают через выгрузку.

Что показывает реестр

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) хотя бы один стандартный способ отдать описания — DCAT, OAI-PMH, CSW, OGC API Records, SDMX или STAC — записан у 9 330 каталогов из 43 614, это 21%. Выгрузка DCAT есть у 5 233. Из них 3 152 отдают сразу и DCAT-US, и DCAT-AP: это в основном сайты ArcGIS Hub, у которых обе ленты включены по умолчанию.

Разметку schema.org реестр почти не записывает — таких адресов в нём 38, — поэтому её долю по реестру считать нельзя. Её видно только на самой странице.

Доля зависит от платформы сильнее, чем от желания владельца. Среди порталов открытых данных DCAT отмечен у 239 из 245 на Socrata, у 348 из 454 на Huwise, у 103 из 161 на DKAN и у 633 из 1 906 на CKAN. У собственных разработок — у 24 из 1 144. Расширение ckanext-dcat ставит треть порталов на CKAN, хотя без него каталог не попадает в сбор. Запросы — в приложении о реестре.

Расширения и их риски

У CKAN в ядре выгрузки описаний нет. Всю работу делает ckanext-dcat. Его ведёт команда CKAN, лицензия AGPL, последний выпуск — 2.4.4 от 11 августа 2026 года. В документации поддержка заявлена для CKAN 2.10 и 2.11, автоматические тесты идут и на 2.12. Пока поддержку 2.12 не объявили, работу расширения на этой версии проверяют на своём стенде. Методика оценки расширения — в главе об экосистемах расширений.

Внутри одного расширения несколько плагинов, и каждый включают отдельно:

- **dcat** отдаёт каталог в RDF и читает чужие выгрузки;
- **structured_data** ставит разметку schema.org на страницу набора. Без него DCAT есть, а поисковик каталог не видит;
- **croissant** отдаёт описание для машинного обучения. Площадки в 2026 году отдают Croissant 1.0, хотя вышла версия 1.1.

Профиль по умолчанию — DCAT-AP 3. Рядом лежат профили DCAT-AP 2, DCAT-US 3, HealthDCAT-AP и schema.org. Национальный профиль пишут своим плагином поверх готового: так устроен DCAT-AP_IT. Риск здесь не в заброшенности расширения, а в настройке. Портал с расширением «для галочки» отдаёт профиль по умолчанию и не проходит проверку своего национального портала.

Отдельное расширение, ckanext-spatial, поднимает CSW через pycsw, но только для записей, собранных с других геопорталов, а не для всего каталога CKAN.

У Dataverse выгрузка DCAT — тоже расширение, и оно моложе. Экспортёр DCAT 3 от сообщества GDCC, выпуск 0.8.7 от 31 августа 2026 года, делает DCAT 3 и нидерландский профиль DCAT-AP-NL 3.0. Он рассчитан на Dataverse 6, лучше всего на 6.2 и новее, и сам помечен как прототип. Ставить его стоит тем, кто уже знает, какой профиль от них ждут.

Как ломается: файл открывается, профиль не тот

Проверка 13 октября 2026 года. Три выгрузки прогнаны через SHACL-валидатор Европейской комиссии, профиль DCAT-AP 3.0.1, уровень базовых правил.

data.europa.eu, набор о качестве воздуха в Бристоле. Выгрузка Turtle прошла без ошибок. Это ожидаемо: портал сам собирает описания по DCAT-AP и сам их отдает. Зато страница набора в браузере без скриптов пуста, и разметки schema.org в ней нет. Каталог образцово открыт для сборщика и закрыт для поисковика, который читает только HTML.

data.gov.gr, ряд наблюдений метеостанции. Каталог на CKAN 2.11.3 с ckanext-dcat, профиль DCAT-AP 3, в описании стоят и признак набора высокой ценности, и ссылка на регламент. Валидатор нашёл две ошибки. Формат файла записан текстом «CSV», а профиль ждёт ссылку на тип из словаря. У контрольной суммы файла не указан алгоритм. Страница при этом размечена: в HTML есть блок schema.org. Выгрузка и разметка сделаны, а профиль формально не проходит из-за двух полей.

data.gouv.fr, набор о субсидиях. Файл Turtle открывается, но валидатор отмечает формат «xls», записанный текстом. На полной проверке добавляются издатель и частота обновления: они записаны ссылками, которые не объявлены объектами нужного типа. В HTML страницы разметки нет совсем.

Ещё четыре поломки встречаются часто.

- **Профиль по умолчанию.** Расширение включено, но отдаёт DCAT-AP 2 там, где ждут DCAT-AP 3, или DCAT-US 1.1 там, где уже принимают 3.0.
- **Выгрузка без страницы.** Сборщик каталог забирает, поисковик — нет. Так устроены udata и ArcGIS Hub.
- **Страница без выгрузки.** Разметка есть, а файла со всеми описаниями нет. Поисковик наборы найдёт, национальный портал — нет.
- **Чужое в выгрузке.** Лента ArcGIS Hub описывает всё содержимое сайта: на opendata.dc.gov в ней 1 871 запись, из которых наборами поиск считает 1 476. Сборщик, который верит ленте, опубликует у себя карты и приложения как наборы.

Публикация без портала

Без портала описания не появляются сами: их кладёт автор. Хватает двух файлов.

Разметка на странице набора. Блок JSON-LD со словарём schema.org: название, описание, автор, издатель, дата, лицензия и список файлов. Готовый шаблон есть в книге. Его вставляют в шаблон страницы один раз, и он заполняется из полей набора.

Файл со всеми описаниями. Для небольшого каталога это `data.json` по схеме DCAT-US 1.1: её понимают и сборщики data.gov, и агрегаторы. JKAN собирает такой файл из страниц сам, но шаблон заполняет короткий список полей и не закрывает обязательные поля схемы. Перед сдачей сборщику файл проверяют, см. главу о статических сайтах. На своём статическом сайте его собирают скриптом при каждой сборке, иначе он устареет в день публикации второго набора.

Если набор архивируют в Zenodo, описание в DataCite туда уходит вместе с DOI, и отдельный файл DCAT для этого набора не нужен. Как устроить страницу и регистрацию в каталоге, разобрано в главе «Как сделать публикацию находимой».

Как проверить живой портал

Ручная проверка — четыре вопроса к живому каталогу:

1. Откройте страницу набора и найдите в исходном коде блок `application/ld+json` с типом Dataset. Либо вставьте адрес страницы в проверку расширенных результатов Google.
2. Найдите файл со всеми описаниями и посмотрите, какой профиль он заявляет. У DCAT-US это поле `conformsTo`.
3. Сравните числа. Наборов в выгрузке должно быть столько же, сколько в поиске на сайте.
4. Прогоните файл через валидатор нужного профиля. Для DCAT-AP это SHACL-валидатор Европейской комиссии: берите проверку базовых правил той версии, которую ждёт ваш сборщик, сейчас это `dcatap.3_0_1_base`.

Команды

Подставьте адрес портала и адрес страницы одного набора. Команды проверяют разметку, типичные адреса выгрузки и заявленный профиль.

```
PORTAL=https://data.gov.gr
PAGE=https://data.gov.gr/dataset/emy-station-189

# Разметка schema.org на странице набора
curl -sL "$PAGE" | grep -o '"@type": *"Dataset"' | head -1

# Файл со всеми описаниями: DCAT по страницам и data.json
curl -sI "$PORTAL/catalog.ttl" | head -1
curl -sI "$PORTAL/data.json" | head -1

# Какой профиль заявляет файл
curl -s "$PORTAL/data.json" | head -c 400
```

Дальше — адреса конкретных платформ, проверенные 13 октября 2026 года:

```
# CKAN с ckanext-dcat: описание одного набора в Turtle и число наборов в выгрузке
curl -sL "https://data.gov.gr/dataset/emy-station-189.ttl" | head -20
curl -s "https://data.gov.gr/catalog.ttl" | grep -E "hydra:(totalItems|lastPage)"

# udata: файл Turtle по набору и общая выгрузка
curl -sI "https://www.data.gouv.fr/api/1/datasets/slug-набора/rdf.ttl" | head -1

# Huwise: весь каталог одним файлом
curl -sI "https://opendata.paris.fr/api/explore/v2.1/catalog/exports/dcat" | head -1

# ArcGIS Hub: какая лента включена и сколько в ней записей
for V in dcat-us/1.1 dcat-us/3.0 dcat-ap/2.1.1 dcat-ap/3.0.0; do
  echo -n "$V "; curl -s "https://opendata.dc.gov/api/feed/$V.json" | jq '.dataset | length'
done

# InvenioRDM: одна запись в DCAT и в schema.org
curl -s -H "Accept: application/dcat+xml" "https://data.caltech.edu/api/records/
идентификатор" | head
curl -s -H "Accept: application/ld+json" "https://data.caltech.edu/api/records/идентификатор"
| head

# GeoNetwork 4.4: какие форматы отдаёт CSW
curl -s "https://sdi.eea.europa.eu/catalogue/srv/eng/csw?
service=CSW&version=2.0.2&request=GetCapabilities" \
  | grep -o 'dcat[^<]*'

# Научный репозиторий: OAI-PMH
curl -s "https://dataverse.harvard.edu/oai?verb=Identify" | grep repositoryName

# Dataverse: встроенная разметка Croissant и список форматов выгрузки
curl -s "https://borealisdata.ca/api/info/exportFormats"
```

Хороший результат: блок Dataset в HTML, файл выгрузки открывается без входа, в нём заявлен нужный профиль, а валидатор не находит ошибок базового уровня. Плохой: адрес отвечает, но профиль старый, форматы записаны текстом вместо ссылок на словарь, а на странице набора разметки нет. Сводный список — в чек-листах.

Что из этого следует

- Сначала узнайте, кто заберёт ваши описания и какой профиль он ждёт. Выгрузка «в DCAT вообще» этого требования не закрывает.
- Ставьте и выгрузку, и разметку. Первая кормит сборщика, вторая — поисковик, и платформы закрывают их по отдельности.
- Проверяйте выгрузку валидатором нужной версии профиля, а не только ответом сервера. Формат файла текстом вместо ссылки на словарь — типичная причина отказа.
- На CKAN всё это делает одно расширение, ckanext-dcat, но профиль и разметку включают явно. Расширение без настройки отдаёт профиль по умолчанию.
- Пишете каталог сами — заложите в задание и файл выгрузки в профиле страны, и блок schema.org в шаблоне страницы. Ни то ни другое само не появится.

Источники

- DCAT 3, рекомендация W3C

- DCAT-AP 3.0.1, рекомендация SEMIC от 27 октября 2025 года; черновик 3.0.2 от 17 июня 2026 года; профили GeoDCAT-AP, StatDCAT-AP, DCAT-AP HVD
- DCAT-US 3.0 и схема DCAT-US 1.1
- Разметка Dataset в документации Google — schema.org и DCAT, обязательные поля, фильтр по Croissant
- Документация ckanext-dcat: профили, разметка для Google Dataset Search, выпуск 2.4.4 от 11 августа 2026 года; ckanext-spatial: CSW
- Экспорт метаданных Dataverse, находимость и Croissant — встроен с версии 6.10; экспортёр DCAT 3 от GDCC, выпуск 0.8.7 от 31 августа 2026 года
- Сериализаторы записей InvenioRDM — DCAT, DataCite, Dublin Core, schema.org
- GeoNetwork 4.4 и поисковики — JSON-LD schema.org с Croissant; форматы DCAT в GeoNetwork 4.4
- Каталог pycsw в коде GeoNode и форматы вывода pycsw
- ArcGIS Hub: ленты каталога; ленты DCAT-US 3.0 и DCAT-AP 3.0.0 на opendata.dc.gov — проверка 13 октября 2026 года
- SHACL-валидатор DCAT-AP Европейской комиссии — типы проверок для DCAT-AP 3.0.1, 3.0.0, 2.1.1 и DCAT-AP HVD
- Живые проверки 13 октября 2026 года: data.gov.gr (CKAN 2.11.3, набор emy-station-189), data.gouv.fr (udata 17.12.1), opendata.paris.fr, data.economie.gouv.fr и data.bs.ch (Huwise), data.cityofnewyork.us (Socrata), opendata.dc.gov (ArcGIS Hub), data.europa.eu (набор о качестве воздуха в Бристоле), каталог EEA (GeoNetwork 4.4), каталог Ifremer Sextant (GeoNetwork 4.4.13), GeoNode в Никарагуа и Боливии, библиотека микроданных Всемирного банка (NADA), Borealis (Dataverse 6.8.4), CaltechDATA (InvenioRDM), opendata.scot (JKAN), выгрузка DCAT установки IPT в Новой Зеландии
- И. Бегтин, Почему данные по России так непросто найти, ноябрь 2024, и Поисковые системы по данным как глобальные системы обнаружения данных, май 2022 — какие порталы попадают в поиск по данным и почему
- Распространённость выгрузок — запрос к выгрузке реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. Запросы приведены в приложении
- Стандарты метаданных, находимость, полнота метаданных

Полнота и детальность метаданных

Описание набора решает две задачи: по нему набор находят и по нему решают, можно ли им пользоваться. Отсюда главный практический вывод главы. Короткая обязательная часть, которую заполняет каждый публикатор и проверяет программа. Предметные поля — только там, где без них данные нельзя понять. Словарь полей — у каждой таблицы. И никаких значений-заглушек: пустое поле видно сразу, а заглушка выглядит как заполненное и обманывает и людей, и программы.

Глава разбирает, что требуют стандарты и чего они не требуют, как платформы позволяют добавить свои поля и описать колонки таблиц, и как за полчаса проверить, насколько честно заполнены описания на живом портале. Проверка в конце главы сделана на двух национальных порталах — греческом и французском — и показывает, как формально полное описание оказывается пустым.

Что это

Три свойства описания часто смешивают.

Полнота — какая доля нужных полей заполнена. Её считают по наборам: у скольких наборов из ста есть лицензия, период, издатель.

Детальность — насколько подробна сама схема: только название и описание или ещё период, территория, методика, список колонок. Детальность задаёт платформа и владелец каталога, полноту обеспечивают публикаторы.

Достоверность — правда ли то, что записано в поле. Период «1900–2099» у таблицы за один год, описание, которое повторяет название, территория «вся страна» у данных одного района. Формально поле заполнено, по существу — нет.

Платформа может быть подробной и при этом пустой: двадцать обязательных полей, в которых стоит прочерк. И наоборот: короткая схема, заполненная честно, полезнее длинной анкеты. Крайний случай — каталог вычислительной химии ioChem-BD на DSpace: у его наборов название часто просто числовой код, например «000112758». Химик, который знает, что ищет, такой набор откроет по ссылке из статьи. Поиск по каталогу и поисковики данных его не найдут.

Зачем нужна

Чтобы найти. Поиск и фильтры работают только по заполненным полям. Набор без периода не попадёт в фильтр «2020–2024», набор без темы — в раздел «Транспорт». Поисковики данных строже: Google принимает разметку набора, только если у неё есть название и описание длиной от 50 до 5000 символов.

Чтобы решить, можно ли брать. Без лицензии права пользователя не определены, даже если файл скачивается: что можно делать с данными, зависит от закона страны и от того, охраняются ли они вообще. Осторожный пользователь такой набор не возьмёт. Без издателя

и контакта некому задать вопрос и некому сообщить об ошибке.

Чтобы понять, что внутри. Колонка `sum_2` без словаря полей открыта формально и закрыта по смыслу. В 2021 году Росводресурсы открыли систему «Водные данные» с паспортами наборов, API и электронной подписью файлов. Описания структуры полей в паспортах не было, а в одном поле реестра водопользователей вместе лежали название владельца, адрес и ИНН. Данные есть, документации к ним нет.

Чтобы набор пережил своих авторов. Через десять лет никто не вспомнит, что значит код `7` в колонке «статус». Архивы и научные репозитории поэтому описывают подробнее порталов открытых данных: они работают на будущего читателя, а не на сегодняшний отчёт. Пять моделей работы с метаданными — научная, архивная, общественно-государственная, технологическая и коммерческая — разобраны в главе о классификации каталогов. От модели зависит, какая глубина описания для каталога нормальна.

Чтобы сборщик не гадал. Агрегатор, который собирает описания со многих каталогов, при бедном описании набора вынужден подставлять свойства самого каталога: если весь каталог посвящён здравоохранению, его наборы получают тему «здоровье». Так поисковик Dateno ищет слои с серверов ArcGIS и GeoServer, у которых своего описания почти нет. Это лучше, чем ничего, но точность такого поиска — точность догадки.

Четыре уровня описания

Минимум, без которого набор нельзя использовать:

- название и описание человеческим языком, не повторяющее название;
- кто опубликовал и как с ним связаться;
- лицензия;
- дата последнего обновления данных — не правки карточки;
- постоянный адрес.

Этого хватает, чтобы найти набор и понять, можно ли его брать.

Рекомендуемый уровень добавляет то, без чего набор трудно отфильтровать и трудно обновлять:

- охват по времени и территории;
- как часто обновляется;
- тема из общего справочника;
- формат и размер файла;
- язык.

Предметный уровень нужен там, где без него данные нельзя интерпретировать:

- статистика — методика, единица измерения, разрывы ряда;
- геоданные — система координат, масштаб, тип объекта;
- наука — авторы с ORCID, финансирование, связь со статьёй;
- микроданные — выборка, вопросник, уровни доступа.

Описание содержимого — словарь полей: что означает каждая колонка, какой у неё тип и единица, что значит пустая ячейка. Это метаданные уже не о наборе, а о его содержимом, и устроены они отдельно. О них — раздел «Словарь полей» ниже.

Чем отличаются даты данных от дат карточки, разобрано в главе «Анатомия каталога». Коротко: если платформа их не различает, дата «обновлено» говорит о правке описания, а не о свежести данных.

Что требуют стандарты и чего не требуют

Обязательные поля стандарта — это минимум для обмена между программами, а не минимум пользы. Самое неожиданное в таблице ниже — где оказалась лицензия.

Стандарт	Обязательно	Важное, но не обязательное
DCAT-AP 3.0.1, профиль порталов ЕС	у набора — название и описание; у способа получения — адрес доступа	у набора 7 рекомендуемых свойств: контакт, способ получения, территория, ключевые слова, издатель, период, тема. Ещё 27 необязательных, среди них дата выпуска, дата изменения и частота обновления. Лицензия — рекомендуемое свойство способа получения, а не набора
Схема DataCite, по ней регистрируют DOI	идентификатор, автор, название, издатель, год публикации, тип ресурса	описание, тема, даты, территория — рекомендуемые; лицензия, язык, формат, размер, версия, финансирование — необязательные
schema.org/Dataset по требованиям Google	название, описание от 50 до 5000 символов	лицензия, автор, период, территория, измеряемые переменные, идентификатор — рекомендуемые

Из этого следуют два вывода.

Первый: свой список обязательного составляют от пользователя, а не копируют из стандарта. Лицензия, дата обновления и период для пользователя обязательны, хотя ни один из трёх стандартов их так не называет.

Второй: в DCAT-AP лицензия живёт у файла. Платформа, построенная по профилю, хранит её там же, и проверка «есть ли лицензия у набора» на ней даст неверный ответ. Пример — в разделе о проверке живого портала.

Если значение действительно неизвестно, честнее сказать об этом, чем выдумать. В схеме DataCite для этого есть стандартные коды: `:unkn` — неизвестно, `:none` — значения нет, `:null` — пусто намеренно. Цена такой честности видна в примере из документации DataCite: ссылка на набор, где неизвестно почти всё, выглядит как `:unkn (2020). :none. :null.` Зато её никто не примет за полноценную.

Варианты реализации

Платформы позволяют настроить схему описания пятью способами. От способа зависит, кто может поменять поля и сколько это стоит при обновлении.

Готовая схема. Поля заданы платформой, поменять их нельзя или можно только правкой кода. Так устроены лёгкие каталоги вроде JKAN и статический сайт: схему пишет автор шаблона, проверки полей нет. У ArcGIS Online и сайтов ArcGIS Hub своих полей нет, но администратор выбирает для организации один стиль описания из готовых стандартов: ISO

19139, ISO 19115-3, INSPIRE, FGDC, североамериканский профиль ISO 19115 или Dublin Core+. У организаций, созданных после июля 2025 года, по умолчанию стоит Dublin Core+, у более старых — ISO 19139.

Схема из файла. Администратор описывает поля в файле, платформа строит по нему форму и проверку.

- В CKAN это делает расширение ckanext-scheming. Схема пишется в JSON или YAML и задаёт поля набора, ресурса, организации и группы: обязательность, список допустимых значений, готовые заготовки полей, функции проверки. С CKAN 2.12 можно не требовать обязательные поля у черновика, пока его не опубликовали.
- В DKAN схема набора — файл JSON Schema, по умолчанию построенный на DCAT-US. Свой файл кладут вместо стандартного, и всё, что попадает в каталог, проверяется по нему. Документация предупреждает: менять схему, когда в каталоге уже есть наборы, нужно осторожно, несовместимая правка ломает проверку старых записей.
- В DSpace поля берутся из реестра схем метаданных (по умолчанию Dublin Core), а формы ввода описаны в файле `submission-forms.xml`. Поле можно показывать только для определённого типа записи, например только для набора данных. После правки форм сервер перезапускают.

Блоки и профили. Платформа поставляет готовые наборы полей под отрасли, их включают и дополняют своими.

- В Dataverse семь поддерживаемых блоков: цитирование, геоданные, социальные и гуманитарные науки, астрономия, науки о жизни, журналы, 3D-объекты. Есть и экспериментальные: CodeMeta для программ, вычислительные процессы, архивные сведения, метки Local Contexts для знаний коренных народов. Администратор коллекции сам решает, какие поля в ней обязательны. Свой блок описывают в файле TSV и загружают через API. Документация отдельно напоминает: новые поля нужно добавить и в схему поискового индекса Solr, иначе по ним не будет поиска.
- В GeoNetwork 4.4 в ядре схемы ISO 19139, ISO 19115-3, Dublin Core и ISO 19110 — каталог объектов, то есть словарь полей для геоданных. Национальные профили ставятся плагинами схем: швейцарский, канадский HNAP, нидерландские, итальянский RNDT, британский GEMINI, а также DCAT-AP.
- InvenioRDM хранит описание по схеме DataCite. Обязательны тип ресурса, название, дата публикации и авторы. Готовые дополнительные поля — сведения о журнале, издании, диссертации, конференции и CodeMeta для программ. Свои поля задают в настройках сайта и раскладывают по разделам формы.

Шаблоны в облачном сервисе. Свои поля заводит администратор в веб-интерфейсе, без программиста.

- У Huwise (бывший OpenDataSoft) базовый шаблон «Standard» не меняется и не удаляется. Рядом подключаются шаблоны для совместимости — DCAT, INSPIRE — и свои: администратор создаёт поля с типом, допустимыми значениями и правилами и включает перевод по каждому полю. Отдельные служебные шаблоны видны только в административной части.

- У Socrata (Tyler Data & Insights) свои поля собираются в наборы полей. У каждого поля своя обязательность — без обязательного поля набор не опубликовать — и видимость: всем или только сотрудникам. К полю можно приписать проверку на языке запросов SoQL, в том числе зависящую от других полей, а инструмент аудита показывает, какие наборы правилам не соответствуют. В файл `data.json` для data.gov попадают только поля с заранее известными именами из набора «Common Core».

Шаблон как часть стандарта. Полный стандарт для публикатора слишком велик: в DDI или ISO 19115 сотни элементов. Удачное решение — у редактора метаданных Всемирного банка, с которым работают каталоги микроданных на NADA. Для каждого типа данных в нём два вида готовых шаблонов, оба неизменяемые: полный, со всеми элементами стандарта, и рекомендуемый, только с главными. Организация копирует шаблон и правит копию: убирает лишнее, меняет подписи, делает поля обязательными. Стандарт и выгрузка остаются общими, а форма у каждого издателя своя. Тот же приём применим на любой платформе: схема одна, а формы для разных публикаторов разные.

Словарь полей

Словарь полей описывает колонки таблицы: название, понятная подпись, смысл, тип, единица измерения, допустимые значения, обозначение пропуска. Без него таблица открыта формально и непонятна по существу. У набора метеостанции Парнита на data.gov.gr в хранилище таблиц все 57 колонок записаны как текст, включая температуру и давление, и ни у одной нет описания. Карточка целиком разобрана в главе «Анатомия каталога».

Машиночитаемых форм у словаря несколько, и выбирают ту, что принята в отрасли:

Форма	Где применяется
Table Schema из семейства Frictionless, обычно внутри файла `datapackage.json`	открытые данные, публикация в Git-репозитории, проверка файлов программой
CSV on the Web (CSVW), рекомендация W3C 2015 года	таблицы со связанными данными; Google в бета-режиме читает словарь таблицы в этой форме прямо из разметки страницы
переменные DDI	микроданные: у каждой переменной текст вопроса, коды ответов, частоты
ISO 19110, каталог объектов	геоданные: типы объектов слоя и их атрибуты
поля и наборы записей Croissant	наборы для машинного обучения
структура SDMX	показатели: разрезы, справочники, атрибуты

Словарь полезен вдвойне, если по нему проверяют сам файл. Тогда «число» в словаре означает, что в колонке действительно числа, а не текст «н/д». Во Франции это устроено на уровне страны: в каталоге schema.data.gouv.fr в октябре 2026 года 106 общих схем, из них 69 в Table Schema. Ресурс набора на data.gouv.fr ссылается на схему, и портал проверяет файл по ней. Общая схема для одной темы — например, для зарядных станций электромобилей — позволяет склеить файлы сотен издателей в одну таблицу. Но даже там ссылку на схему ставят редко: в выборке из тысячи наборов она была у 36.

Связь набора со стандартом в описании DCAT записывают полем `dct:conformsTo`: оно говорит, какой схеме или отраслевому стандарту соответствует файл. Сами отраслевые стандарты данных — OCDS, GTFS, Darwin Core и другие — разобраны в главе о том, что публикуют.

Когда подробность вредна

Обязательная схема из пятидесяти полей замедляет публикацию. Поля начинают заполнять формально: копируют описание из соседнего набора, повторяют название в описании, ставят сегодняшнюю дату в «период данных». Хуже всего, когда это делает сама программа: подставляет значение по умолчанию, чтобы форма прошла проверку. Как это выглядит на национальном портале, показано ниже, в проверке data.gov.gr.

Отсюда три правила.

- Лучше короткая обязательная схема и необязательные предметные поля, чем длинная анкета, которую никто не читает.
- У поля, которое публикатор может не знать, не должно быть значения по умолчанию. Пустое поле видно в любой проверке, заглушку не видно ни в одной.
- Обязательность поля проверяют на трёх-пяти реальных публикаторах до запуска. Если половина из них не может его заполнить, поле делают необязательным или помогают заполнять — справочником, подсказкой, автоматическим расчётом из файла.

Свои поля нужны, если отраслевой справочник или закон требует реквизита, которого нет в стандартной схеме: категория набора высокой ценности, правовое основание публикации, код программы финансирования. Они не нужны, если хочется «на всякий случай» сохранить внутренний номер из учётной системы. Обратная крайность тоже встречается: тщательно описывать то, что каталогизировать вовсе не нужно, например каждую служебную выгрузку промежуточного расчёта.

Когда нужна, а когда нет

Ситуация	Что сделать	Чем обычно закрывают
Портал открытых данных в ЕС, который собирает национальный портал	DCAT-AP и национальный профиль; лицензия у каждого файла; темы и периодичность из словарей ЕС	CKAN со scheming и ckanext-dcat, Piveau, Huwise с шаблоном DCAT
Небольшой муниципальный портал	минимум и рекомендуемый уровень; свои поля не заводить; словарь полей у таблиц	готовая схема облачного сервиса или CKAN без scheming
Научный репозиторий	DataCite плюс блок дисциплины; авторы с ORCID, связь со статьёй	Dataverse с отраслевыми блоками, InvenioRDM со своими полями
Геопортал	ISO 19139 или ISO 19115-3, профиль страны, каталог объектов у векторных слоёв	GeoNetwork с плагином профиля, стиль ISO в ArcGIS
Каталог микроданных	DDI с описанием переменных	NADA с редактором метаданных, Dataverse
Каталог показателей	структура SDMX, методика, единицы	см. главу о каталогах показателей

Ситуация	Что сделать	Чем обычно закрывают
Публикация без портала	файл README, `datapackage.json` со словарём, разметка schema.org на странице	шаблоны этой книги, проверка в CI

Свою схему полей стоит заводить, когда её требует закон, профиль верхнего портала или отрасль. Её не стоит заводить ради полноты анкеты: каждое поле — это работа публикатора на каждом наборе и работа администратора при каждом обновлении платформы.

Коротко, как выглядит обоснование по шаблону карточки решения:

- **Контекст.** Региональный портал на CKAN, национальный профиль требует два поля, которых нет в ядре: категорию набора высокой ценности и правовое основание.
- **Варианты.** Писать значения в свободные дополнительные поля (`extras`); поставить ckanext-scheming; перейти на облачный сервис с шаблонами.
- **Критерии.** Проверка значений при вводе, выгрузка в DCAT-AP, стоимость обновлений.
- **Решение.** ckanext-scheming со схемой в YAML и ckanext-dcat с профилем страны.
- **Почему.** Свободные поля не проверяются и попадают в выгрузку под случайными именами. Переход на облачный сервис ради двух полей дороже, чем одно расширение, которое ведёт сообщество CKAN.
- **Когда пересмотреть.** Если профиль страны начнут поставлять готовым расширением или если обязательных своих полей станет больше десятка.

Как это сделано в платформах

Пометки — ядро, расширение, нет, не проверено, как в сводной таблице. Сверено с документацией платформ 12 октября 2026 года.

Платформа	Свои поля описания	Словарь полей таблицы	Что важно знать
CKAN 2.12	расширение ckanext-scheming: схема в JSON или YAML	ядро: словарь у таблиц в хранилище DataStore — подпись, описание, тип колонки; с 2.11 ещё Table Designer — таблица со схемой и проверкой значений	несколько языков описания — расширение ckanext-fluent; проверка файлов по Table Schema — расширение ckanext-validation, см. ниже
DKAN	ядро: замена файла JSON Schema	не проверено	по умолчанию схема DCAT-US; менять её на живом каталоге рискованно
Huwise (OpenDataSoft)	ядро: свои шаблоны полей, служебные шаблоны	ядро: подпись, описание и один из 12 типов у каждого поля; вкладка «Схема» показывает долю заполнения каждой колонки	базовый шаблон не меняется
Socrata	ядро: наборы своих полей с обязательностью и проверкой SoQL	ядро: название, описание и тип колонки, отдаются в `/api/views/ID.json`	в `data.json` уходят только поля набора «Common Core»
ArcGIS Online и Hub	нет: один стиль из готовых стандартов на организацию	ядро: у размещённого слоя описание поля, тип значения, списки допустимых значений	описание слоя и описание элемента в каталоге не синхронизируются
Dataverse	ядро: отраслевые блоки и свои блоки в TSV	ядро: у загруженных таблиц (SPSS, Stata, R, Excel, CSV) переменные описываются в DDI	новые поля нужно добавить в схему индекса Solr

Платформа	Свои поля описания	Словарь полей таблицы	Что важно знать
InvenioRDM	ядро: свои поля в настройках	нет	основа — схема DataCite
DSpace	ядро: реестр схем и `submission-forms.xml`	нет	правка форм требует перезапуска
GeoNetwork 4.4	ядро: ISO 19139, ISO 19115-3, Dublin Core; профиль страны — плагин схемы	ядро: каталог объектов ISO 19110	плагин нужен ровно под вашу версию GeoNetwork
NADA	не проверено	ядро: переменные DDI с частотами	шаблоны полей — в отдельном редакторе метаданных Всемирного банка
JKAN и статический сайт	нет: поля задаёт шаблон	нет: словарь кладут файлом рядом	проверку пишет автор сайта

Гибкая схема почти всегда означает сопровождение: кто-то должен решить, какое поле обязательно, и не дать схеме расползтись. Чем глубже свои поля встроены в платформу, тем дороже каждое её обновление.

Расширения и их риски

Свои поля и словарь — та область, где цена расширений видна лучше всего. Методика оценки расширения разобрана в главе об экосистемах расширений, здесь — то, что нашлось при сверке.

- **ckanext-scheming и ckanext-fluent** ведёт сообщество CKAN, автоматические тесты идут на CKAN 2.10, 2.11 и 2.12. Это обычная часть взрослого портала на CKAN, а не экзотика. Подробности — в главе про CKAN.
- **ckanext-validation** проверяет загруженные таблицы по Table Schema с помощью Frictionless. В основном репозитории Frictionless Data в октябре 2026 года совместимость заявлена только с CKAN 2.9, для 2.10 и 2.11 в таблице совместимости стоит «нет», и тесты идут только на 2.9. Живая ветка — форк правительства Квинсленда qld-gov-au/ckanext-validation: его тесты идут на CKAN 2.10, 2.11 и 2.12. Это типичная судьба расширения: исходный автор отошёл, а работающую версию поддерживает один из пользователей. Перед установкой проверяют не название, а репозиторий и ветку.
- **Плагины схем GeoNetwork** привязаны к версии ядра: на каждую линейку GeoNetwork у плагина своя ветка. У одного плагина бывают копии в разных местах, и они расходятся. Британский профиль GEMINI 2.3 ведёт компания Astun Technology, в её репозитории есть ветка для GeoNetwork 4.4. В общем собрании плагинов metadata101 лежит копия, которая не обновлялась с февраля 2023 года и доходит только до 4.2.
- **Свои блоки Dataverse и своя схема DKAN** — не расширения, а настройка ядра. Но и их проверяют при каждом обновлении: блок Dataverse требует согласованной схемы индекса, а несовместимая правка схемы DKAN ломает проверку старых записей.

Публикация без портала

Без портала полнота описания целиком на авторе: никакая форма не потребует лицензию. Зато и детальность ничем не ограничена. Хватает трёх файлов рядом с данными:

- `README.md` — что это, за какой период, кто отвечает, как обновляется, лицензия;
- `datapackage.json` — список файлов и словарь полей каждой таблицы по Table Schema;
- разметка schema.org/Dataset на странице набора, если у набора есть свой сайт.

Готовые README, datapackage.json и разметка schema.org лежат в шаблонах книги. Проверку можно поручить программе: команда `frictionless validate datapackage.json`, запущенная в CI репозитория при каждом изменении, сверит файлы со словарём и не даст выложить таблицу, где в колонке чисел появился текст. Как устроить такой репозиторий, разобрано в главе о Git-репозиториях.

Как проверить живой портал

Ручная проверка — по-прежнему лучший первый шаг:

1. Откройте десять случайных наборов из разных организаций, не самые свежие.
2. Для каждого отметьте, заполнены ли лицензия, издатель, контакт, дата обновления, период и территория.
3. Сравните описание с названием: не повторяет ли оно его.
4. Посмотрите, есть ли словарь полей у табличного набора и совпадают ли типы колонок с содержимым.
5. Проверьте, похожи ли значения на правду: период — на содержимое файла, территория — на издателя.

Пятый шаг важнее остальных. Валидатор профиля и оценка качества метаданных на data.europa.eu (MQA) проверяют, что поле есть, а не что в нём правда. Заглушка проходит обе проверки.

Две проверки на выборке

На большом портале проверку делают программой. 12 октября 2026 года через API взята тысяча наборов на каждом из двух национальных порталов: десять блоков по сто наборов, равномерно по всему каталогу. Числа — снимок на этот день.

data.gov.gr — национальный портал Греции на CKAN 2.11 со схемой по DCAT-AP, всего около 22,7 тысячи наборов. В выборку попали наборы 118 организаций.

Поле	Похоже на заполненное	Заполнено по существу
Период	97%	у 938 наборов из 1000 период «1900-01-01 — 2099-12-31», одинаковый у министерства, государственного архива и десятков муниципалитетов. Его подставила программа, а не издатель
Территория	96%	у 60% название территории — пустая строка, ещё у 35% — вся Греция
Контакт и издатель	24% и 31%	22% и 28%: у части записей все поля контакта — пустые строки

Поле	Похоже на заполненное	Заполнено по существу
Описание	94%	у 31% описание короче 50 символов, у 22% оно дословно повторяет название
Лицензия	у 32% наборов, если искать её у набора	у 99% наборов, если искать её у файлов, как велит DCAT-AP
Частота обновления	36%	у 32% выборки значение «не планируется», настоящая частота — только у 4%

Заглушка периода уходит и в выгрузку DCAT, которую собирает европейский портал: в описании набора стоят `dcat:startDate "1900-01-01"` и `dcat:endDate "2099-12-31"`, а рядом пустое английское название. Сборщик и оценка качества увидят период заполненным у почти всех наборов страны.

data.gouv.fr — национальный портал Франции на udata, всего около 74,4 тысячи наборов, 78% выборки собраны с других каталогов. У udata в ядре есть своя оценка качества описания: девять равных критериев, от лицензии и периода до доступности файлов. API отдаёт её по каждому набору, на сайте она видна рядом с карточкой. Медиана по выборке — 5 из 9.

- Лицензия «не указана» у 19% наборов. При этом встроенная оценка засчитала лицензию у 98%: по исходному коду udata критерий выполнен, если поле лицензии вообще заполнено, а «не указана» — тоже значение.
- Период указан у 23% наборов, частота обновления «неизвестна» у 70%.
- Территория: уровень детальности (коммуна, департамент) указан у 63%, а сами территории — у 14%.
- Описание длиннее 100 символов, как требует встроенная оценка, у 73%.

Оба портала — хорошие, живые и большие. Именно поэтому на них видно главное: формальная полнота и полнота по существу расходятся в разы, а встроенная оценка и внешний валидатор этого расхождения не замечают.

Команды

Для портала на CKAN. Команда берёт 500 наборов и считает долю коротких описаний, описаний, повторяющих название, лицензию у набора и у файлов, и два самых частых значения периода. Имя поля периода зависит от схемы портала: на data.gov.gr это `temporal_coverage`, на другом портале посмотрите его в ответе `package_show`.

```
PORTAL=https://data.gov.gr
curl -s "$PORTAL/api/3/action/package_search?rows=500&start=5000" \
  | jq '.result.results as $r | ($r | length) as $n | {
      datasets: $n,
      short_description: ([$r[] | select((.notes // "") | length < 50)] | length / $n),
      description_is_title: ([$r[] | select((.notes // "" | gsub("^\\s+|\\s+$"; "")) ==
.title)] | length / $n),
      license_on_dataset: ([$r[] | select((.license_id // "") != "")] | length / $n),
      license_on_files: ([$r[] | select(any(.resources[]; (.license // "") != ""))] | length
/ $n),
      top_periods: ([$r[] | .temporal_coverage | tostring] | group_by(.) | map({value: .[0],
n: length}) | sort_by(-.n) | .[0:2])
    }'
```

Самое полезное здесь — самые частые значения. Если одно значение стоит у большинства наборов разных организаций, это заглушка. Тот же приём годится для территории, частоты обновления и контакта.

Для портала на udata. Команда сравнивает лицензию со встроенной оценкой и считает период и частоту на одной странице выдачи:

```
curl -s "https://www.data.gouv.fr/api/1/datasets/?page_size=100&page=200" \
  | jq '.data as $d | ($d | length) as $n | {
      datasets: $n,
      license_not_specified: ([$d[] | select(.license == "notspecified")] | length / $n),
      score_counts_license: ([$d[] | select(.quality.license)] | length / $n),
      frequency_unknown: ([$d[] | select(.frequency == "unknown")] | length / $n),
      with_period: ([$d[] | select(.temporal_coverage != null)] | length / $n)
    }'
```

Запускайте команды несколько раз с разными `start` или `page`, чтобы выборка не состояла из одной организации. Для портала на Socrata словарь полей и даты отдаются по адресу `/api/views/ID-набора.json`: там у каждой колонки название, тип и описание, а дата обновления данных записана отдельно от даты правки описания. Как превратить такие доли в метрики и что с ними делать дальше — в главе о метриках качества.

Что из этого следует

- Составьте свой список обязательных полей от пользователя: лицензия, издатель, контакт, дата обновления данных, период. Не ограничивайтесь обязательными полями стандарта.
- Уберите значения по умолчанию у полей, которые публикатор может не знать. Пустое поле лучше заглушки.
- Требуйте словарь полей у каждой таблицы и, где можно, проверяйте файл по нему.
- Свои поля заводите под закон, профиль или отрасль, а не «на всякий случай». Каждое поле стоит работы на каждом наборе и при каждом обновлении платформы.
- Расширения для схем проверяйте по репозиторию и ветке под свою версию ядра, а не по названию.
- Считайте полноту по существу: самые частые значения, повторы названия в описании, правдоподобие периода. Встроенным оценкам и валидаторам верьте только в том, что поле есть.

Источники

- DCAT-AP 3.0.1 — обязательные, рекомендуемые и необязательные свойства набора и способа получения; проверено 12 октября 2026 года
- Схема метаданных DataCite — обязательные и рекомендуемые свойства, коды для неизвестных значений
- Разметка Dataset в документации Google — обязательные свойства, длина описания, словарь таблиц в CSVW
- CKAN 2.12: ckanext-scheming, словарь полей DataStore, Table Designer, журнал изменений 2.11

- ckanext-validation: репозиторий Frictionless Data — таблица совместимости и тесты только на CKAN 2.9; форк qld-gov-au — тесты на CKAN 2.10–2.12; проверено 12 октября 2026 года
- DKAN: замена схемы набора
- Dataverse: поддерживаемые блоки метаданных, настройка метаданных, загрузка табличных файлов и описание переменных
- InvenioRDM: справочник метаданных, свои поля
- DSpace: формы ввода
- GeoNetwork: схемы ядра в выпуске 4.4.12 от 8 июля 2026 года, плагины metadata101, GEMINI 2.3 у Astun Technology
- Huwise: шаблоны метаданных, схема набора, вкладка «Схема»
- Socrata: свои поля описания, поля для data.json
- ArcGIS: стили метаданных, описание полей слоя
- Всемирный банк: редактор метаданных и шаблоны
- Table Schema, CSV on the Web, каталог схем schema.data.gouv.fr
- udata: расчёт оценки качества набора
- Выборки: API data.gov.gr и API data.gouv.fr, по 1000 наборов, 12 октября 2026 года; ioChem-BD, набор 000112758
- И. Бегтин, Telegram: модели работы с метаданными, 10.05.2021; «Водные данные» Росводресурсов, 21.08.2021; ioChem-BD, 10.01.2025; свойства каталога переходят на наборы, 15.01.2025; оценка качества начинается с полноты метаданных, 01.08.2025
- Анатомия каталога, экспорт метаданных, метрики качества, стандарты метаданных

API и массовая выгрузка

Человеку достаточно сайта. Программе нужен устойчивый способ спросить каталог: какие есть наборы, что в описании, где файл, что изменилось со вчерашнего дня. Это и есть программный интерфейс, API.

Главный практический вывод главы. Любому каталогу нужны две вещи: способ забрать все описания сразу и прямые ссылки на файлы, и то и другое без регистрации. API поиска и API к содержимому таблиц добавляют по размеру каталога и по задачам пользователей. Лимиты на API честны, только если рядом есть полная выгрузка: иначе каждый, кому нужна вся база, выкачивает её по кусочку, и это дорого и ему, и серверу.

Глава разбирает четыре задачи, которые обычно называют одним словом «API», варианты их решения, условия доступа и лимиты, интерфейсы, о которых владелец сайта не знает, и то, как это устроено в платформах. В конце — проверка десяти живых каталогов на разных платформах 12 октября 2026 года и команды, которыми её можно повторить.

Что это

API (программный интерфейс) — адрес, по которому программа задаёт каталогу вопрос и получает ответ, обычно в JSON или XML. Важно не само слово «API», а то, на какой вопрос интерфейс отвечает. Таких вопросов четыре.

1. **Найти и прочитать описание.** Поиск по тексту и фильтрам, карточка одного набора. Пример — Action API у CKAN, Search API у Dataverse.
2. **Забрать все описания сразу.** Один файл или протокол, который отдаёт каталог целиком: файл `data.json`, выгрузка DCAT, протоколы OAI-PMH и CSW. Это нужно сборщикам: вышестоящему порталу, поисковику данных, архиву.
3. **Получить сами данные.** Ссылка на файл, запрос к таблице, картографический сервис, статистический API.
4. **Узнать, что изменилось.** Лента RSS или Atom, выборка «изменено после даты», файлы изменений. Чтобы не скачивать каталог заново каждый день.

Ещё три термина понадобятся дальше.

Массовая выгрузка — возможность скачать всю базу или весь набор файлами, без запросов по одной записи. Регламент ЕС о наборах высокой ценности определяет её так: скачивание всего набора одним или несколькими пакетами.

Постраничная выдача — ответ, разбитый на страницы: по 100 наборов, по 25 записей. Программа идёт со страницы на страницу, пока не дойдёт до последней. Так устроены почти все API поиска и многие выгрузки DCAT.

Лента изменений — список последних добавленных и изменённых записей в формате RSS или Atom. Её читают программы-подписчики и сборщики.

Не путайте API каталога с каталогом API. Здесь API — дверь в каталог наборов. Каталог API — отдельный тип сайта, где сами интерфейсы и есть то, что описывают; он разобран в главе о других типах каталогов.

Зачем нужна

Без программного входа каталог виден только тем, кто открыл сайт в браузере. Вот кто остаётся за бортом:

- **вышестоящий портал и агрегаторы.** Национальный портал собирает региональные, data.europa.eu собирает национальные, поисковики данных собирают всех. Им нужны все описания и изменения за день. Как устроены такие цепочки — в главе о федерациях и сборе;
- **поисковые системы.** Им достаточно карты сайта и разметки на страницах, о ней — в главе об экспорте метаданных;
- **разработчики и аналитики.** Приложение с расписанием автобусов или ежедневный отчёт не будут вручную скачивать файл;
- **архивы.** Веб-архив сохраняет страницы, но плохо сохраняет данные за API. Когда каталог закрывают, спасти удаётся то, что можно выгрузить целиком. Подробнее — в главе об архивации;
- **модели и ИИ-агенты.** Серверы MCP у порталов данных работают поверх того же API. Если API нет, ассистенту остаётся читать страницы сайта. Подробнее — в главе о трендах;
- **сам владелец.** Многие сайты каталогов построены поверх своего API, а интерфейс для людей — лишь один из его клиентов.

Для части данных API и массовая выгрузка обязательны по закону. Регламент ЕС 2023/138 о наборах высокой ценности (геоданные, наблюдения Земли и окружающая среда, метеорология, статистика, компании, мобильность) требует от органов власти:

- отдавать наборы через API, а там, где это указано в приложении к регламенту, ещё и массовой выгрузкой. Приложение требует и то и другое для всех шести категорий, с отдельными исключениями: например, данные численных моделей прогноза погоды — только через API;
- опубликовать условия использования API и показатели качества обслуживания: производительность, пропускную способность, доступность;
- приложить документацию API в открытом формате, понятном и человеку, и программе, — на практике это описание OpenAPI;
- назначить контакт для вопросов об API.

Эти четыре требования — хороший список и для тех, на кого регламент не распространяется.

Варианты реализации

Все описания сразу

Способ	Как устроен	Сильная сторона	Слабое место	Где встречается
Один файл	весь каталог в одном файле: `data.json` по DCAT-US, файл DCAT, таблица CSV	забирается одной командой, понятен без документации	растёт вместе с каталогом; на сотнях тысяч наборов становится тяжёлым	Socrata, ArcGIS Hub, Huwise, DKAN, статические сайты
Выгрузка DCAT по страницам	тот же DCAT, разбитый на страницы со ссылками на следующую и последнюю	не нагружает сервер, работает на любом размере	надо пройти все страницы; если каталог меняется во время обхода, наборы могут потеряться или повториться	CKAN с ckanext-dcat, udata
Протокол сбора	OAI-PMH у научных репозиториев, CSW и OGC API Records у геопорталов	стандарт, есть выборка по дате изменения	медленно: сотни тысяч записей — тысячи запросов; в OAI-PMH нет понятия файла	Dataverse, InvenioRDM, DSpace, GeoNetwork
Регулярная выгрузка каталога файлами	раз в сутки сервер сам собирает весь каталог в файлы и выкладывает их как обычный набор	самый дешёвый способ для сервера и для сборщика	данные на момент выгрузки, а не на сейчас	data.gouv.fr: описания наборов, файлов, сервисов, организаций в CSV
Дамп базы	копия базы данных целиком: SQL, RDF, архив JSON	всё, включая связи, которых нет в описаниях	нужен навык, чтобы развернуть; формат зависит от платформы	Сенат Франции, Europeana
Карта сайта	список адресов страниц	её понимают все поисковые системы	только адреса, без описаний	почти все платформы, в CKAN — расширением

Выгрузка всех описаний нужна почти всегда. Это минимум, по которому каталог могут собрать другие. Карта сайта её не заменяет: поисковику она подскажет, какие страницы обойти, а сборщику каталога ничего не скажет о лицензии и файлах.

Поиск по API выгрузку тоже не заменяет. У API поиска почти всегда есть предел на одну страницу и на глубину выдачи. В CKAN по умолчанию за один запрос отдаётся не больше 1000 наборов, у Zenodo без ключа — 25 записей на страницу и не дальше 10 000-й записи. Пройти весь каталог через поиск иногда можно, иногда нет, и в любом случае это хрупко.

Поиск и карточка

Единого стандарта API поиска наборов нет. У каждой платформы свой: Action API у CKAN, API у udata, Explore API у Huwise, SODA и Discovery API у Socrata, Search API у Dataverse, REST API у InvenioRDM и DSpace. Ближе всего к общему стандарту OGC API Records: его отдают сайты ArcGIS Hub и GeoNetwork с отдельным сервисом. Поэтому вышестоящие порталы собирают описания по DCAT, а не через API каждой платформы.

API поиска нужен, когда наборов сотни и больше, когда каталог входит в федерацию или когда поверх него строят свой сайт. Для двадцати наборов на статическом сайте хватает одного файла описаний и карты сайта.

Сами данные

- **Прямая ссылка на файл.** Самый простой и самый живучий способ. Её понимают человек, сборщик и архив.
- **API к таблице.** Строки с фильтром, сортировкой и постраничной выдачей. В CKAN это DataStore: по умолчанию до 32 000 строк за запрос, а SQL-запросы выключены, пока администратор их не включит. В DKAN, Huwise и Socrata API к таблицам — в ядре; Huwise отдаёт набор и целиком файлом, в том числе в Parquet.
- **API к каждому файлу без загрузчика.** Каталог сам разбирает опубликованный табличный файл и через несколько минут отдаёт к нему API. Так работает табличный API data.gouv.fr (пока в бете). Предел — 100 МБ для CSV, 50 МБ для XLS и Parquet, 12,5 МБ для XLSX. У каждого файла есть профиль колонок, постраничная выдача с фильтрами и описание запросов в Swagger. Код сервиса открыт под лицензией MIT. Hugging Face так же переводит наборы в Parquet и даёт к ним консоль SQL прямо в браузере. Подробнее — в главе о передовых каталогах.
- **Картографические сервисы.** WMS отдаёт картинку слоя, WFS и OGC API Features — сами объекты, ArcGIS REST — и то и другое. Подробнее — в главе о геопорталах.
- **Статистический API.** SDMX, PxWebApi, JSON-stat: запрос к ряду или кубу по кодам разрезов. Подробнее — в главе о каталогах показателей.
- **Данные в облаке.** Адрес бакета S3, каталог STAC, файл Parquet, который читают по частям прямо по сети.
- **Выборка и расчёт на сервере.** Каталог отдаёт не файл, а нужный кусок или результат. ERDDAP принимает в адресе выборку по месту и времени и формат ответа. Сервер openEO выполняет цепочку операций над снимками и отдаёт итог. Climate Data Store показывает готовый код запроса к выборке, собранной в форме.
- **Поток.** Данные в реальном времени: положение транспорта в GTFS-RT, показания датчиков через SensorThings API.

API к содержимому таблиц нужен, когда данные большие или обновляются часто, а скачивать весь файл ради одной строки неудобно. Для архива годового отчёта достаточно файла. Автоматический API наследует ошибки исходного файла: кодировку, лишние строки над заголовком, текст в числовой колонке. Без словаря колонок он отдаёт их быстрее, но не точнее.

Что изменилось

- **Лента RSS или Atom.** Есть в ядре CKAN, udata, Huwise, ArcGIS Hub и DSpace. Длина у неё разная: в CKAN по умолчанию 20 последних записей, на data.gouv.fr — 50, а у Huwise и ArcGIS Hub в ленту попадает весь каталог. Короткая лента годится для подписки, но не для полного сбора: если сборщик пропустил день, часть изменений уйдёт за её край.
- **Выборка «изменено после даты».** Параметр `from` в OAI-PMH, `modified_since` в выгрузке DCAT у ckanext-dcat, фильтр по полю `metadata_modified` в поиске CKAN, `pastDays` в PxWebApi 2. Это главный инструмент сборщика: полная выгрузка раз в месяц и выборка изменений каждый день.
- **Журнал действий.** В CKAN — плагин activity из поставки ядра: кто, когда и что поменял.

- **Файлы изменений.** Для больших баз рядом с полной выгрузкой кладут разницу с прошлой версией. Так делают адресный реестр ФИАС и OpenSanctions.

Уведомления об изменениях нужны сборщикам больших каталогов. Маленькому порталу достаточно честной даты изменения в выгрузке. Честной — значит даты изменения данных, а не правки карточки: чем они отличаются, разобрано в главе «Анатомия каталога».

Массовая выгрузка — не роскошь

Если данные можно получить только запросами к API, каждый, кому нужна вся база, выкачивает её по кусочку. Это долго для пользователя и дорого для сервера. Цифры показывают, насколько это лишнее. Европейский центральный банк отдаёт все ряды массовой выгрузкой: в 2024 году это были 108 файлов CSV на 93 ГБ, которые после перевода в Parquet заняли 664 МБ. Всемирный банк выкладывает архивом основную базу World Development Indicators, но в январе 2025 года для остальных показателей data.worldbank.org массовой выгрузки не было. 12% показателей, скачанных через API, — это 3382 файла JSON на 22 ГБ, а в Parquet те же данные заняли 44 МБ. Довод «данные слишком большие, поэтому только API» обычно значит, что их не пробовали сжать.

Ограничения API разумны, только если рядом есть полная выгрузка. Zenodo в 2024 году ограничил выдачу одного поискового запроса десятью тысячами записей, а 12 октября 2026 года наборов данных в нём 726 858. Файла со всеми описаниями у Zenodo нет. Забрать всё можно только по OAI-PMH — запрос за запросом, по полсотни записей. Это работает, но для каталога такого размера медленно, и большинство пользователей об этом пути не знают.

Хорошие образцы обратного:

- **data.gouv.fr** каждую ночь выкладывает весь каталог файлами CSV как обычный набор данных. 12 октября 2026 года описания наборов заняли 276 МБ, описания файлов — 1,1 ГБ, отдельно лежат сервисы данных, организации, примеры использования и источники сбора;
- **Сенат Франции** выкладывает свои базы дампами для PostgreSQL в архивах ZIP: например, все поправки к законопроектам с 2001 года;
- **адресный реестр ФИАС** в 2023 году публиковал полные выгрузки до 36 ГБ в сжатом виде и отдельные файлы изменений;
- **статистика сельского хозяйства США (NASS)** отдаёт всю базу Quick Stats несколькими сжатыми файлами и рядом даёт API для точечных запросов. 12 октября 2026 года последние файлы были от 10 октября;
- **OpenSanctions** выкладывает заранее собранные выгрузки и файлы изменений в нескольких форматах. Технически это образец. Юридически — нет: бесплатно данные можно использовать только в некоммерческих целях.

Что отличает хорошую выгрузку:

- **постоянный адрес последней версии** и архив прошлых по датам;
- **размер и дата рядом со ссылкой**, чтобы пользователь знал, что скачивает;
- **сжатый формат**: CSV в gzip или Parquet, а не несжатый XML. Как выбрать формат и сжатие — в главе о форматах;

- **описания вместе с данными.** У ЕЦБ в 2024 году массово отдавались значения рядов, а описания показателей приходилось извлекать из самих файлов. Обычно бывает наоборот, и это так же плохо;
- **файлы изменений** для баз, которые целиком весят гигабайты;
- **лицензия у самой выгрузки**, а не только на странице набора.

Условия доступа и лимиты

Метаданные должны читаться без регистрации, без ключа и без капчи. Иначе автоматический сбор либо невозможен, либо ломается при каждом обновлении защиты. Для данных ключ допустим, если он выдаётся сразу, бесплатно и без ручного одобрения: так владелец видит, кто создаёт нагрузку, и может связаться с ним.

Лимиты нормальны, если они опубликованы и предсказуемы. Хороший пример — PxWebApi 2 у статистики Швеции: адрес `/api/v2/config` сам сообщает программе, что за один запрос отдаётся не больше 150 000 ячеек, а запросов — не больше 30 за 10 секунд. Плохой пример — лимит, о котором узнаёшь по ошибке на 401-й странице выдачи.

Что закрывает каталог от программ, даже когда API формально есть:

- **проверка от роботов** на входе: Cloudflare, AWS WAF. Страница «Just a moment…» вместо ответа ломает любой сборщик. Примеры — в главах о научных репозиториях и о каталогах микроданных;
- **геоблокировка**: сайт не отвечает на запросы из других стран;
- **обязательная регистрация** ради защиты от ботов ИИ;
- **временные ссылки** на файлы, которые перестают работать через полчаса.

Подробнее эти случаи разобраны в антипаттерне «Дверь закрыта для программ».

Ещё одно требование, о котором забывают, — устойчивость. Интерфейс, на который опираются другие, нельзя менять молча. Старую версию держат рядом с новой, о закрытии объявляют заранее, адреса не ломают. Когда в апреле 2026 года data.gov ушёл с CKAN на свою разработку, адреса страниц наборов сохранились, а Action API CKAN 12 октября 2026 года отвечал ошибкой 404. Всё, что было построено на старом API, перестало работать в один день.

Неявные интерфейсы

Многие сайты отдают данные программам, даже если владелец об этом не написал. Причин три. API задумали, но не задокументировали. Сайт сам устроен как страница, которая получает данные скриптом из API, и этот API можно вызвать напрямую. Или API есть у платформы по умолчанию, а владелец о нём не знает: у WordPress это `/wp-json/`, у SharePoint — `/_api/`, у научных репозиториев DSpace — REST API и OAI-PMH, у серверов ArcGIS — ответ в JSON на любой адрес с параметром `?f=json`, у каталогов на Blacklight — тот же адрес с заголовком `Accept: application/json`.

Найти такой интерфейс помогают три приёма: проверить адреса по умолчанию для платформы сайта, открыть в браузере инструменты разработчика и посмотреть на вкладке «Сеть», какие запросы страница делает сама, и поискать в поисковике документацию по

фразе «REST API» на домене сайта.

Для пользователя такой интерфейс — находка: иногда это единственный способ получить данные с геопортала или из реестра. Для владельца — повод задокументировать то, что уже есть. Опираться на недокументированный интерфейс в постоянной работе рискованно: при обновлении сайта он меняется без предупреждения. Если данные нужны надолго, выгрузку из такого API сохраняют у себя; для этого есть готовые программы, например apibackuper.

Данные без прямой ссылки

Каталог, привыкший к модели «набор — это файлы по ссылкам», встречает и другие способы доступа.

- **Облачные каталоги.** Каталог Microsoft Planetary Computer отдаёт часть данных ссылками вида `abfs://…` во внутреннее хранилище Azure, реестр открытых данных Amazon — адресами бакетов S3, каталог Google Earth Engine — фрагментом кода для своей платформы. Данные открыты, но удобнее всего ими пользоваться внутри облака владельца.
- **«Сначала код» в машинном обучении.** Kaggle, Hugging Face, каталоги TensorFlow и PyTorch вместо ссылки на файл показывают строку кода, которая загружает набор библиотекой. Для специалиста это удобно, для общего поисковика данных — набор без файлов.
- **Только по запросу.** Часть статистических порталов отдаёт таблицу только в ответ на POST-запрос с выбранными параметрами, без постоянного адреса.

Если вы публикуете так, дайте рядом хотя бы одну обычную ссылку на файл или выгрузку: её поймут и человек, и сборщик, и архив.

Когда нужна, а когда нет

Каталог	Обязательно	Стоит добавить	Можно не делать
Двадцать наборов на статическом сайте или в Git-репозитории	файл со всеми описаниями (`data.json` или DCAT), карта сайта, прямые ссылки на файлы	разметку schema.org на страницах	API поиска, API к таблицам
Муниципальный портал, который собирает национальный	выгрузку DCAT в профиле страны без регистрации, прямые ссылки на файлы	ленту изменений или выборку по дате	свой API к таблицам, если файлы небольшие
Национальный портал, в том числе с наборами высокой ценности ЕС	API поиска, выгрузку всех описаний, выборку изменений; для наборов высокой ценности — API и массовую выгрузку, условия использования, документацию, контакт	ежедневную выгрузку каталога файлами	—
Научный репозиторий	OAI-PMH, REST API, прямые ссылки на файлы	выгрузку всех описаний файлом, если записей сотни тысяч	API к содержимому таблиц
Геопортал	CSW или OGC API Records, сервисы WMS и WFS или OGC API Features, скачивание слоя целиком	DCAT через CSW для портала открытых данных	—

Каталог	Обязательно	Стоит добавить	Можно не делать
Каталог показателей	API к рядам (SDMX, PxWebApi, JSON-stat), список всех таблиц с датами обновления	массовую выгрузку всей базы	—
Большая реестровая база: адреса, компании, санкции	полную выгрузку и файлы изменений	API для запросов по одной записи	—
Набор для машинного обучения	прямую ссылку на файлы, лучше в Parquet, и описание Croissant	строку кода для библиотеки как дополнение, а не замену	—

Коротко, как выглядит обоснование по шаблону карточки решения:

- **Контекст.** Статистическая служба на PxWeb, около 3000 таблиц. Аналитики и поисковик данных просят «дать всё сразу», сервер API в дни публикаций перегружен.
- **Варианты.** Оставить только PxWebApi; поднять отдельный сервис SDMX; добавить ночную выгрузку всех таблиц файлами.
- **Критерии.** Сколько запросов нужно, чтобы забрать всю базу; нагрузка на сервер; требования международной отчётности; стоимость сопровождения.
- **Решение.** PxWebApi 2 для точечных запросов и ночная выгрузка всех таблиц в CSV и Parquet с описью: код таблицы, название, дата обновления, размер.
- **Почему.** Через API всю базу забирают минимум одним запросом на таблицу, а большие таблицы упираются в лимит ячеек и требуют нескольких запросов. Один набор файлов в сутки дешевле серверу, чем тысячи запросов от каждого пользователя. Выгрузку пишут скриптом поверх того же API: в ядре PxWeb её нет.
- **Когда пересмотреть.** Если международные партнёры потребуют SDMX или если выгрузку почти не скачивают.

Как это сделано в платформах

Пометки — ядро, расширение, нет, не проверено, как в сводной таблице. Сверено с документацией и живыми установками 12 октября 2026 года.

Платформа	Поиск описаний	Все описания сразу	Данные	Что изменилось
CKAN 2.12	ядро: Action API, до 1000 наборов за запрос	расширение ckanext-dcat: DCAT по страницам, по умолчанию 100 наборов на странице; карта сайта — тоже расширение	ядро: файлы и DataStore; SQL-запросы выключены по умолчанию; заливка CSV — расширение XLoader	ядро: лента Atom, фильтр по дате изменения в поиске, журнал действий (плагин activity из поставки); в ckanext-dcat — параметр `modified_since`
udata 17	ядро	ядро: DCAT по страницам; ночная выгрузка каталога в CSV — задача ядра, включается настройкой	ядро: файлы	ядро: лента Atom, сортировка по дате обновления
Huwise (OpenDataSoft)	ядро: Explore API	ядро: каталог одним файлом в DCAT, `data.json` и CSV	ядро: API к записям, выгрузка набора в CSV, JSON, Parquet и другие форматы	ядро: лента RSS, фильтр по дате изменения
Socrata	ядро: Discovery API	ядро: `data.json` одним файлом	ядро: SODA	не проверено

Платформа	Поиск описаний	Все описания сразу	Данные	Что изменилось
ArcGIS Hub	ядро: поиск по OGC API Records	ядро: DCAT-US 1.1 и DCAT-AP 2.1.1 одним файлом; DCAT-US 3.0 и DCAT-AP 3.0.0 включаются отдельно	ядро: слои ArcGIS и файлы	ядро: лента RSS 2.0
DKAN	ядро: API метаданных	ядро: `data.json` по DCAT-US	ядро: хранилище таблиц с SQL-запросами	не проверено
Dataverse	ядро: Search API	ядро: OAI-PMH; описание набора — в DDI, DataCite, schema.org и других форматах	ядро: Data Access API	ядро: OAI-PMH с выборкой по дате
InvenioRDM	ядро: REST API	ядро: OAI-PMH, в том числе в DataCite и DCAT	ядро: файлы	ядро: OAI-PMH с выборкой по дате
DSpace 7 и новее	ядро: REST API	ядро: OAI-PMH, карта сайта	ядро: файлы	ядро: ленты Atom и RSS по любому поисковому запросу через OpenSearch, OAI-PMH с выборкой по дате
GeoNetwork 4.4	ядро: свой API поиска	ядро: CSW (в том числе в DCAT-AP), OAI-PMH, карта сайта; OGC API Records — отдельный микросервис проекта	ядро: ссылки на сервисы и файлы	ядро: OAI-PMH с выборкой по дате
PxWeb (PxWebApi 2)	ядро	ядро: список всех таблиц с датами обновления	ядро: JSON-stat 2, CSV, PX, XLSX; SDMX нет	ядро: таблицы, изменённые за последние N дней
NADA	ядро: свой API	ядро: свой API и описание DDI у каждого исследования; OAI-PMH и DCAT нет	ядро: файлы по уровням доступа	не проверено
JKAN, статический сайт	нет	файл `data.json`, если его собрали	ссылки на файлы	лента изменений репозитория, если сайт собран из Git: у GitHub это `commits/ветка.atom`

Три вещи в таблице неочевидны.

У CKAN выгрузка всех описаний — не ядро, а расширение ckanext-dcat. На взрослом портале оно стоит почти всегда, но без него сборщику остаётся только Action API. Карты сайта в ядре CKAN тоже нет: data.gov.gr на неё отвечает ошибкой 404.

У GeoNetwork 4 часть прежних интерфейсов убрана: OpenSearch недоступен с версии 4.0, отдельная служба DCAT `rdf.search` заменена выводом DCAT через CSW. OGC API Records ставится рядом отдельным сервисом. В каталоге Европейского агентства по окружающей среде на GeoNetwork 4.4 этот адрес 12 октября 2026 года не отвечал, а CSW и OAI-PMH работали.

У PxWeb второго поколения есть то, чего не было в первом: список всех таблиц с датой обновления и выборка изменённых за N дней. Массовой выгрузки самих данных в ядре по-прежнему нет.

Что показывает реестр

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) хотя бы один программный вход записан у 28 471 каталога из 43 614, это 65%. Стандартный способ забрать все описания — DCAT, OAI-PMH, CSW или OGC API Records — у 8 894, это 20%. Остальные отдают то, что умеет их платформа: интерфейс ArcGIS, API CKAN без расширения, свой API облачного сервиса.

Главное различие — не между типами каталогов, а между платформами. У каталогов на готовых платформах программный вход записан у 76%, у собственных разработок — у 10% (697 из 7 034), а стандартная выгрузка описаний — у 58 из 7 034, меньше 1%. Отчасти это свойство самого реестра: он находит интерфейсы по известным адресам платформ, а у своей разработки адреса заранее неизвестны. Но и сборщик сталкивается с тем же. Если вы пишете каталог сами, стандартную выгрузку придётся сделать самим, иначе каталог не соберёт никто.

По типам разброс тоже большой. Программный вход есть у 70% геопорталов, 69% порталов открытых данных и 63% научных репозиториев, но только у 37% каталогов показателей и 23% ML-каталогов. Ленты RSS и Atom реестр у порталов на CKAN почти не записывает, хотя они в ядре, поэтому долю лент по реестру не считаем. Запросы — в приложении о реестре.

Как проверить живой портал

Документация для разработчиков на самом сайте — первый признак. Дальше пять шагов:

1. **Определите платформу.** От неё зависит, по каким адресам искать: CKAN отвечает на `/api/3/action/status_show`, Dataverse — на `/api/info/version`, DSpace 7 и новее — на `/server/api`.
2. **Найдите все описания сразу** и дойдите до последней страницы выгрузки.
3. **Сравните числа.** Наборов в выгрузке должно быть столько же, сколько в поиске на сайте. Если поиск говорит, что наборов 10 000, а выгрузка содержит 200, сборщик увидит не тот каталог, который видит человек.
4. **Проверьте выборку изменений**: за последние дни должно что-то находиться, если портал живой.
5. **Повторите без браузера.** Запрос из командной строки, без входа, лучше из другой сети. Если в ответ приходит проверка от роботов или ошибка 403, программы этот каталог не увидят.

Десять каталогов 12 октября 2026 года

Каталог	Платформа	Все описания сразу	Что изменилось	Что заметно
data.gov.gr	CKAN 2.11.3 с ckanext-dcat	DCAT по страницам: 455 страниц по 50 наборов, всего 22 736 — столько же, сколько в поиске	лента Atom; 1328 наборов изменены с 1 октября — и по поиску, и по выгрузке DCAT	карты сайта нет

Каталог	Платформа	Все описания сразу	Что изменилось	Что заметно
data.gouv.fr	udata 17.12.1	DCAT по страницам: 744 страницы по 100, всего 74 375; ночная выгрузка каталога в CSV	лента Atom, сортировка по обновлению	ночная выгрузка заменяет 744 запроса к DCAT одним скачиванием; описания файлов — отдельный CSV на 1,1 ГБ
opendata.paris.fr	Huwise	один файл DCAT на 14 МБ, `data.json` и CSV на 490 наборов	лента RSS; 96 наборов изменены с 1 октября	каждый набор выгружается и в Parquet
data.cityofnewyork.us	Socrata	`data.json` одним файлом, 4,8 МБ, 2711 записей	не проверялось	общий поиск по всем порталам Socrata работает без ключа
opendata.dc.gov	ArcGIS Hub	DCAT-US 1.1 и DCAT-AP 2.1.1 одним файлом, 1871 запись; ленты DCAT-US 3.0 и DCAT-AP 3.0.0 на том же сайте тоже отвечают	лента RSS 2.0	в выгрузку попадают и документы, карты, приложения: наборов в поиске 1476 из 1864 элементов
Harvard Dataverse	Dataverse 6.10.1	OAI-PMH; в поиске 305 251 набор	OAI-PMH с выборкой по дате	страница набора отвечает роботу проверкой AWS WAF, а API и OAI-PMH работают
Zenodo	InvenioRDM	только OAI-PMH, в том числе в DCAT и DataCite; 726 858 наборов	OAI-PMH с выборкой по дате	поиск: без ключа по 25 записей, не дальше 10 000-й
DSpace@MIT	DSpace 8.2	OAI-PMH, карта сайта	лента Atom через OpenSearch	REST API на `/server/api`
каталог EEA	GeoNetwork 4.4	CSW: 3429 записей; OAI-PMH; карта сайта	OAI-PMH с выборкой по дате	OGC API Records не развёрнут
статистика Швеции	PxWebApi 2.3.2	список всех 2814 таблиц	53 таблицы изменены за 7 дней	лимиты опубликованы в `/api/v2/config`

Общий вывод проверки: у всех десяти каталогов есть способ забрать все описания, но устроен он по-разному, и универсального сборщика нет. Ни один из проверенных порталов не отвечает по адресу `/.well-known/api-catalog`, который стандарт RFC 9727 предлагает для описания всех API сайта.

Команды

Для портала на CKAN. Команды проверяют версию, выгрузку всех описаний, совпадение чисел и выборку изменений. Подставьте адрес портала и дату.

```
PORTAL=https://data.gov.gr

# Версия CKAN
curl -s "$PORTAL/api/3/action/status_show" | jq '.result.ckan_version'

# Все описания: сколько наборов и где последняя страница выгрузки DCAT
curl -s "$PORTAL/catalog.ttl" | grep -E "hydra:(totalItems|lastPage)"

# Столько же наборов должно быть в поиске
curl -s "$PORTAL/api/3/action/package_search?rows=0" | jq '.result.count'

# Что изменилось с 1 октября: в поиске и в выгрузке DCAT
curl -s -G "$PORTAL/api/3/action/package_search" --data-urlencode "rows=0" \
```

```
  --data-urlencode "fq=metadata_modified:[2026-10-01T00:00:00Z TO *]" | jq '.result.count'
curl -s "$PORTAL/catalog.ttl?modified_since=2026-10-01" | grep "hydra:totalItems"

# Лента изменений
curl -sI "$PORTAL/feeds/dataset.atom" | head -1
```

Для других платформ:

```
# udata: выгрузка DCAT и ночные файлы со всем каталогом
curl -s "https://www.data.gouv.fr/api/1/site/catalog.rdf" | grep -o "hydra:totalItems[^<]*"
curl -s "https://www.data.gouv.fr/api/1/datasets/catalogue-des-donnees-de-data-gouv-fr/" \
  | jq -r '.resources[] | "\(.title)\t\(.filesize)"'

# Huwise: весь каталог одним файлом и наборы, изменённые после даты
curl -sI "https://opendata.paris.fr/api/explore/v2.1/catalog/exports/dcat" | head -1
curl -s -G "https://opendata.paris.fr/api/explore/v2.1/catalog/datasets" \
  --data-urlencode "limit=0" --data-urlencode 'where=modified >= "2026-10-01"' | jq
'.total_count'

# Socrata и ArcGIS Hub: файл со всеми описаниями
curl -s "https://data.cityofnewyork.us/data.json" | jq '.dataset | length'
curl -s "https://opendata.dc.gov/api/feed/dcat-us/1.1.json" | jq '.dataset | length'

# Научный репозиторий: OAI-PMH и записи, изменённые после даты
curl -s "https://dataverse.harvard.edu/oai?verb=Identify" | grep -o "<repositoryName>[^<]*"
curl -s "https://zenodo.org/oai2d?verb=ListIdentifiers&metadataPrefix=oai_dc&from=2026-10-11"
\
  | grep -o "<header" | wc -l

# Геопортал: сколько записей отдаёт CSW
curl -s "https://sdi.eea.europa.eu/catalogue/srv/eng/csw?
service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=hits" \
  | grep -o 'numberOfRecordsMatched="[0-9]*"'

# PxWebApi 2: все таблицы, изменённые за неделю, и лимиты
SCB=https://statistikdatabasen.scb.se/api/v2
curl -s "$SCB/tables?lang=en&pageSize=1" | jq '.page.totalElements'
curl -s "$SCB/tables?lang=en&pageSize=1&pastDays=7" | jq '.page.totalElements'
curl -s "$SCB/config" | jq -c '{maxDataCells, maxCallsPerTimeWindow, timeWindow}'

# Предел глубины поиска: 401-я страница по 25 записей у Zenodo даёт ошибку 400
curl -s -o /dev/null -w "%{http_code}\n" "https://zenodo.org/api/records?
type=dataset&size=25&page=401"
```

Неявные интерфейсы платформ. Подставьте домен сайта:

```
curl -s "https://portal.example/wp-json/" | head          # WordPress
curl -s "https://portal.example/_api/web" | head          # SharePoint
curl -s "https://portal.example/server/api" | head        # DSpace 7 и новее
curl -s "https://gis.example/arcgis/rest/services?f=json" # сервер ArcGIS
curl -s -H "Accept: application/json" "https://portal.example/catalog?q=" | head  #
Blacklight и другие
```

Хороший результат: ответ без входа в систему, в нём есть описания наборов, а не только главная страница, а выгрузка доходит до последней страницы. Плохой: капча, обязательная регистрация, ошибка 403 для программ или поиск, который не умеет отдать следующую страницу. Сводный список — в чек-листах.

Что из этого следует

- Сначала сделайте выгрузку всех описаний и прямые ссылки на файлы. Это минимум для любого каталога, от двадцати наборов на статическом сайте до национального портала.
- Метаданные отдавайте без регистрации, ключа и капчи. Проверяйте это из командной строки и из другой сети, а не из своего браузера.
- Дайте сборщикам выборку «изменено после даты». Это дешевле для всех, чем ежедневная полная выгрузка.
- Для больших баз публикуйте полную выгрузку файлами и файлы изменений. Лимиты API без выгрузки — это закрытие данных, а не защита сервера.
- Публикуйте лимиты, условия использования и документацию API, назначьте контакт. Для наборов высокой ценности в ЕС это обязанность, для остальных — хорошая практика.
- Не меняйте интерфейс молча. При переезде платформы сохраните старые адреса или объявите их закрытие заранее.
- Выбирая платформу, проверяйте, что выгрузка описаний и выборка изменений в ядре, а не в расширении. Если пишете каталог сами, заложите стандартную выгрузку в задание: без неё каталог никто не соберёт.

Источники

- Регламент ЕС 2023/138 о наборах высокой ценности — текст на EUR-Lex: статья 2 (определение массовой выгрузки), статья 3 (API, массовая выгрузка, условия использования, качество обслуживания, документация, контакт)
- CKAN 2.12: документация API, настройки — `ckan.search.rows_max`, `ckan.datastore.search.rows_max`, `ckan.datastore.sqlsearch.enabled`, `ckan.feeds.limit`; ckanext-dcat: точки выгрузки RDF — постраничная выгрузка и `modified_since`
- udata: исходный код, задача выгрузки каталога в CSV; выгрузка каталога data.gouv.fr
- Руководство Dataverse по API, REST API WordPress, REST API DSpace 7+, ArcGIS REST API
- GeoNetwork 4.4: интерфейсы, OAI-PMH, OpenSearch убран в 4.0, DCAT через CSW и форматтеры, микросервисы с OGC API Records
- PxWebApi 2 статистики Швеции: список таблиц и настройки с лимитами, версия 2.3.2
- Массовые выгрузки: ЕЦБ, WDI Всемирного банка, NASS Quick Stats, OpenSanctions и условия использования, Сенат Франции, Europeana
- apibackuper — сохранение данных из постраничного API
- Проверки 12 октября 2026 года: выгрузки, ленты и фильтры по дате на data.gov.gr, data.gouv.fr, opendata.paris.fr, data.cityofnewyork.us, opendata.dc.gov; OAI-PMH и поиск Harvard Dataverse, Zenodo, DSpace@MIT; CSW и OAI-PMH каталога EEA; PxWebApi статистики Швеции; ответ 404 у Action API data.gov; адрес `/.well-known/api-catalog` на data.gouv.fr и opendata.dc.gov
- И. Бегтин, Telegram: массовая выгрузка ЕЦБ, 20.08.2024; показатели Всемирного банка, 24.01.2025; ограничения API Zenodo, 06.09.2024; дампы SQL Сената Франции, 15.11.2022; ФИАС, 29.05.2023; NASS, 01.08.2024; OpenSanctions, 23.10.2024; скрытые интерфейсы

платформ, 28.11.2021; три причины недокументированных API, 30.08.2022; как их искать, 26.08.2024; облачные каталоги без прямых ссылок, 27.03.2024; «сначала код» в ML-каталогах, 21.03.2024

- Распространённость интерфейсов — выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. Запросы приведены в приложении.
- data.gouv.fr, API tabulaire (beta) — пределы 100 МБ для CSV, 50 МБ для XLS и Parquet, 12,5 МБ для XLSX, профиль и Swagger у каждого файла; код datagouv/api-tabular, лицензия MIT. Hugging Face, Dataset viewer и SQL Console. ERDDAP, openEO в Copernicus Data Space Ecosystem, Climate Data Store: API. Проверено 15 октября 2026 года

Внутреннее хранилище

Каталог может хранить сами файлы или только описывать их и давать ссылку туда, где файл лежит. От этого выбора зависит, кто отвечает за то, что файл откроется через пять лет, и можно ли быть уверенным, что это тот же самый файл.

Практический вывод главы такой. Научному репозиторию и любому каталогу, на файлы которого ссылаются годами, нужно своё хранилище, и файл после публикации в нём не меняют: исправление — это новая версия. Порталу, который собирает чужие описания, и геопорталу, который показывает слои с картографических серверов, достаточно ссылок, но тогда ссылки нужно регулярно проверять, а у важных наборов держать копию. Таблицу в базе с API заводят только для таблиц, которые фильтруют и запрашивают, а не для каждого файла.

В конце главы — замер на четырёх больших порталах 12 октября 2026 года: какая доля файлов у них загружена, какая ведёт наружу и что происходит со ссылками.

Что это

Под словом «хранилище» прячутся три разные вещи.

Хранилище файлов держит файл целиком: CSV, архив, снимок, PDF с методикой. Пользователь скачивает его по адресу каталога.

Хранилище таблиц — отдельная база, куда загружены строки таблицы. Из неё портал показывает таблицу в браузере, фильтрует её и отдаёт часть строк через API, не заставляя скачивать файл целиком. В CKAN это DataStore, у облачных сервисов — их собственная база записей.

Ссылка — хранилища нет вовсе: каталог хранит только описание и адрес файла или сервиса на чужом сайте.

В описании по DCAT разница видна в полях способа получения. Прямая ссылка на файл записывается в `dcat:downloadURL`, страница или сервис, через которые данные можно получить, — в `dcat:accessURL`. Ссылка на страницу ведомства — это уже не файл, а обещание, что файл где-то там найдётся. Подробнее о частях описания — в главе «Анатомия каталога».

«Внутреннее» не значит «на своём сервере». Почти все крупные порталы держат файлы в облачном объектном хранилище: греческий data.gov.gr — в Microsoft Azure, Humanitarian Data Exchange (HDX) и швейцарский opendata.swiss — в Amazon S3. Объектное хранилище — облачный склад файлов, где каждый файл лежит по своему ключу и отдаётся по HTTP. Пользователь этого не видит: он нажимает «Скачать» на странице портала. У HDX адрес портала переадресует на временную подписанную ссылку S3, которая 12 октября 2026 года жила три минуты. Ссылаться нужно на адрес портала, а не на тот, что оказался в строке браузера после скачивания.

Зачем нужна

Чтобы файл не пропал вместе с чужим сайтом. Если файл лежит в каталоге, он доступен, пока жив каталог. Если каталог только ссылается на сайт ведомства, ссылка ломается при каждой переделке того сайта. При полной выгрузке data.gov.uk в 2023 году около 10% ссылок на файлы не открывались, а к 2026 году, по оценке самой службы, нерабочих ссылок стало больше четверти. Портал собирал ссылки на сайты муниципалитетов, а те меняли платформу или домен. Тот же вопрос в 2026 году разбирала Исследовательская служба Конгресса США: data.gov по закону — единая точка входа, которая ведёт на данные на сайтах ведомств, и в отчёте прямо поставлен вопрос, не должен ли он стать хранилищем, чтобы доступ к данным не пропадал вместе с сайтом ведомства.

Чтобы файл был тем же самым. Статья, отчёт или модель машинного обучения ссылаются на конкретные данные. Если по той же ссылке через год лежит другой файл, вывод уже не проверить. Своё хранилище позволяет посчитать контрольную сумму — короткий «отпечаток» содержимого, который меняется при изменении хотя бы одного байта, — и запретить подмену файла после публикации. С чужим сайтом ни то ни другое не гарантировать.

Чтобы работать с содержимым. Таблицу в базе можно показать в браузере, отфильтровать, отдать частями через API, перевести в другой формат. Над внешней ссылкой это тоже возможно, но только если каталог сам скачает файл, — то есть всё равно заведёт хранилище.

Чтобы проверить файл при загрузке. Кодировка, разделитель, заголовки, соответствие словарю полей проверяются в момент, когда файл приходит в каталог. Если файл живёт на чужом сайте, проверять приходится каждый раз заново, а исправить его каталог не может.

Чтобы проще было сохранить. Архив снимают с одного сайта, а не со ста. Как веб-архивы теряют данные, которые лежат по ссылкам на чужих серверах, — в главе «Архивация и спасение каталогов».

Обратная сторона — обязанности. Своё хранилище — это место, резервные копии, плата за исходящий трафик, удаление файла по требованию (персональные данные, ошибка, авторское право) и ответ на вопрос, кто следит за всем этим через пять лет. Документация DSpace, например, отдельно предупреждает, что при хранении в S3 возможна плата за трафик между хранилищем и сервером, если сервер стоит вне облака Amazon.

Варианты реализации

Только ссылки

Каталог — указатель. Так устроены многие геопорталы: слой живёт на картографическом сервере, каталог хранит адрес сервиса. Так же устроены сборщики: data.europa.eu, поисковики данных, национальные порталы, которые собирают описания ведомств. Чешский data.gov.cz не хранит ни одного файла: 12 октября 2026 года все адреса скачивания примерно 31 тысячи наборов вели на сайты публикаторов. В замере в конце главы у ирландского и швейцарского национальных порталов в выборке нет ни одного загруженного файла.

Это нормальная модель, если источники устойчивы и сами ведут свои данные. Слабое место одно: каталог не может сохранить копию, если источник удалит файл, и узнаёт об этом последним.

Ссылка с проверкой

Каталог хранит ссылку, но регулярно её открывает и записывает результат. Так работает data.gouv.fr: отдельный обходчик hydra проверяет каждый ресурс, а в описании ресурса в API видны дата последней проверки, код ответа сервера, доступность и контрольная сумма. В CKAN то же делает расширение ckanext-archiver: оно скачивает файлы по ссылкам, показывает на странице ресурса, что ссылка сломана, и собирает отчёт о нерабочих ссылках для публикаторов.

Проверка не спасает файл, но превращает «ссылки как-то ломаются» в список, с которым можно работать: какие источники и с какого дня перестали отвечать.

Ссылка с копией

Следующий шаг — хранить копию того, на что ссылаешься. ckanext-archiver предлагает пользователю сохранённую копию, если исходная ссылка сломалась. Обходчик data.gouv.fr идёт дальше: таблицы CSV и Excel, в том числе внешние, он загружает в PostgreSQL и переводит в Parquet, а для этих таблиц работает отдельный табличный API. На внешних ссылках ирландского data.gov.ie в замере тоже лежат копии таблиц в DataStore: портал скачивает файлы Центрального статистического управления и загружает их строки в свою базу.

Копия решает задачу доступа, но порождает вопрос, какая версия главная. Пользователь должен видеть, когда копия снята и совпадает ли она с источником сейчас.

Файлы у каталога

Публикатор загружает файл, каталог его хранит. Технически вариантов три, и выбирают их до запуска: перенос уже загруженных файлов болезненнее, чем выбор в первый день.

- **Диск сервера.** Самое простое и дешёвое для небольшого каталога. Резервные копии, место и перенос — на администраторе.
- **Объектное хранилище** — S3, Azure Blob, Google Cloud Storage или совместимые с S3 системы на своих серверах вроде MinIO и Ceph. Место почти не ограничено, копии делает поставщик, но появляется плата за хранение и скачивание. Пользователь получает файл либо через сервер каталога, либо по переадресации на временную ссылку хранилища, как на HDX.
- **Несколько хранилищ сразу.** Dataverse позволяет выбрать хранилище для каждой коллекции или набора, DSpace — завести несколько пронумерованных хранилищ и переносить файлы между ними, CKAN с версии 2.12 — держать файлы ресурсов, логотипы и вложения расширений в разных местах. Это нужно, когда у разных данных разные требования: большие снимки — в дешёвом облаке, чувствительные файлы — на своём сервере.

Для больших файлов важно, чтобы они не шли через сервер каталога. Dataverse умеет загрузку прямо в S3, минуя сервер, InvenioRDM — загрузку частями, которую можно продолжить после обрыва связи. InvenioRDM умеет и забрать файл по ссылке: сервер сам скачивает его и дальше хранит как свой.

Доверенное внешнее хранилище

Промежуточный вариант для очень больших и чувствительных данных: файл остаётся у владельца, но каталог знает его точный адрес и правила доступа. В Dataverse это «доверенное удалённое хранилище» (trusted remote store): все файлы должны лежать по адресам, начинающимся с заданного базового адреса, а служебные файлы набора — описания, миниатюры — хранятся у самого Dataverse. В InvenioRDM такой файл помечается как удалённый, и документация честно предупреждает: доступность и целостность такого файла платформа не гарантирует.

Таблица в базе

Файл CSV не только лежит, но и загружен в базу. Тогда можно посмотреть первые строки в браузере, запросить нужные через API и не скачивать всё. В CKAN это DataStore — отдельная база PostgreSQL, где каждая таблица привязана к ресурсу. Поиск по умолчанию отдаёт 100 строк, за один запрос — не больше 32 тысяч, запросы на SQL выключены, пока администратор их не включит. Таблицу целиком скачивают по адресу `/datastore/dump/ID-ресурса` в CSV, TSV, JSON или XML.

У облачных сервисов Huwise (бывший OpenDataSoft) и Socrata таблица в базе — основной режим: набор и есть таблица записей, файл получается выгрузкой из неё. Dataverse разбирает загруженные таблицы SPSS, Stata, R, Excel и CSV и описывает их переменные. У каталогов показателей — PxWeb, .Stat Suite — файлов нет вовсе, только база и API к ней.

Таблица в базе удобна для больших и часто обновляемых таблиц и бессмысленна для карты в PDF. Её цена — ещё одна база, которую нужно обновлять, и загрузчик, который ломается на «грязных» файлах: неожиданная кодировка, лишние строки над заголовком, разные типы в одной колонке.

Загрузчик можно вынести из рук издателя. data.gouv.fr, как сказано выше, сам загружает в базу каждый табличный файл, а Hugging Face так же переводит наборы в Parquet. Издатель при этом ничего не делает, но «грязный» файл и тогда даёт «грязный» API. Подробнее — в главе о передовых каталогах.

Смесь

Часть наборов хранится, часть ссылается наружу. Так устроено большинство порталов открытых данных: реестр выгружается файлом в каталог, датчики и карты остаются на сервере ведомства. Смесь честна, если пользователь видит, какой способ у какого ресурса, а каталог проверяет внешние ссылки.

Неизменность, версии и контрольные суммы

Своё хранилище даёт возможность не менять файл после публикации, но не обязывает. Платформы ведут себя по-разному.

- **CKAN в классическом режиме** перезаписывает файл ресурса при обновлении: адрес тот же, содержимое новое, прежнее не сохраняется. Для портала открытых данных, где ресурс «бюджет текущего года» обновляют каждый месяц, это удобно. Для данных, на которые ссылаются в исследованиях, — нет. В CKAN 2.12 появился новый способ работы с файлами: у каждого файла свой идентификатор и контрольная сумма, а изменить содержимое уже созданного файла нельзя, только удалить и создать новый.
- **Zenodo** разрешает поправить файлы в течение 30 дней после публикации. Позже — только новая версия записи со своим DOI, а файлы старой версии остаются доступны, если автор их не закроет. Менять файлы старше 30 дней можно лишь в исключительных случаях: нарушение авторских прав, раскрытие персональных данных.
- **Dataverse** при замене файла в опубликованном наборе создаёт новую версию набора, прежняя остаётся. Для каждого файла считается контрольная сумма, по умолчанию MD5, на выбор SHA-1, SHA-256 или SHA-512.
- **DSpace** хранит контрольную сумму каждого файла и умеет регулярно сверять её с содержимым.

Контрольную сумму полезно отдавать и наружу. В DCAT 3 у способа получения для этого есть свойство `spdx:checksum`, data.gouv.fr показывает сумму в описании ресурса. Тогда пользователь, сборщик и архив могут проверить, что скачали тот же файл, а не исправленный тихо. Как устроены версии и идентификаторы — в главе «Постоянные идентификаторы, DOI и версии».

Размеры и лимиты

У хранилища почти всегда есть предел размера файла, и о нём узнают, когда публикатор приходит с большим архивом.

Где	Предел	Что делать сверх него
CKAN, настройки по умолчанию	10 МиБ на загружаемый файл (`ckan.max_resource_size`)	поднять предел, если хватает места и сети; большие файлы класть в объектное хранилище и ссылаться
Zenodo	50 ГБ и 100 файлов на запись	один раз до 200 ГБ по запросу; дробить набор ради обхода предела правила запрещают
Harvard Dataverse	2,5 ГБ на файл через браузер	крупнее — по договорённости; на своём Dataverse предел задаётся для каждого хранилища
GitHub	100 МиБ на файл, репозиторий лучше держать меньше 1 ГБ	Git LFS, вложения к выпуску, отдельное хранилище
Hugging Face	для открытых наборов место «по возможности», для закрытых — 100 ГБ	см. главу о ML-каталогах

Предел CKAN по умолчанию особенно коварен: 10 МиБ — это небольшая таблица. Портал, где предел не подняли, незаметно превращается в портал ссылок: всё, что крупнее, публикаторы кладут на свой сайт. Лимиты и цены научных репозиториев подробно — в главе о научных репозиториях.

Для данных в терабайты каталог обычно хранит описание, а сами файлы лежат в объектном хранилище, открытом на чтение всем, в форматах, которые читают по частям: Parquet для таблиц, Cloud Optimized GeoTIFF для снимков, Zarr для многомерных массивов. Так устроены Registry of Open Data on AWS и Source Cooperative — о них в главе об облачных хранилищах и хабах.

Когда нужна, а когда нет

Ситуация	Что выбрать	Почему
Научный репозиторий, данные к статьям	свои файлы, без перезаписи, версии, контрольные суммы	статья ссылается на конкретный файл через DOI, подмена ломает проверяемость
Портал организации, у которой данные лежат «на общем диске»	свои файлы	надёжного места у публикатора нет, ссылаться некуда
Национальный или региональный портал, который собирает ведомства и муниципалитеты	ссылки с регулярной проверкой; копии важных наборов	дублировать всё дорого и даёт две расходящиеся копии, но о пропаже файла нужно узнавать первым
Геопортал над картографическими серверами	ссылки на сервисы; файлы — только для выгрузок целиком	слой живёт на сервере, каталог его не заменяет
Каталог показателей	база показателей и API, файлов нет	пользователь берёт срез, а не файл
Большие таблицы, которые фильтруют программы и люди	таблица в базе с API плюс выгрузка целиком	частичный доступ без скачивания всего; выгрузка — для тех, кому нужно всё
Квартальный отчёт из одного файла	только файл	таблица в базе ничего не добавит
Терабайты снимков, моделей, выгрузок	объектное хранилище, описание в каталоге	через сервер каталога такие файлы не пройдут
Чувствительные данные	файл у владельца, в каталоге — описание и порядок доступа	см. главу о доступе и ролях

Коротко, как выглядит обоснование по шаблону карточки решения:

- **Контекст.** Региональный портал на CKAN собирает описания 40 муниципалитетов. Половина ресурсов — ссылки на их сайты, часть сайтов в прошлом году переехала, жалобы на «битые ссылки» приходят ежемесячно.
- **Варианты.** Оставить как есть; перенести все файлы на портал; оставить ссылки, но проверять их и хранить копии.
- **Критерии.** Доля нерабочих ссылок, стоимость места и трафика, кто обновляет данные, риск двух расходящихся копий.
- **Решение.** Ссылки остаются, ckanext-archiver проверяет их каждую неделю и хранит копии; отчёт о нерабочих ссылках раз в месяц уходит публикаторам. Муниципалитетам без своего сайта предложено загружать файлы прямо на портал.
- **Почему.** Перенос всех файлов делает портал ответственным за свежесть чужих данных, а муниципалитеты продолжат обновлять свои сайты, и копии разойдутся. Проверка с копиями дешевле и даёт список проблем по источникам.

- **Когда пересмотреть.** Если доля нерабочих ссылок за полгода не снизится или если портал станет единственным местом публикации по закону.

Как это сделано в платформах

Пометки — ядро, расширение, нет, не проверено, как в сводной таблице. Сверено с документацией и репозиториями 12 октября 2026 года.

Платформа	Файлы у себя	Облачное хранилище	Ссылки наружу	Таблица в базе и API	Файл после публикации
CKAN 2.12	ядро	расширение: в 2.12 — ckanext-file-keeper-cloud, раньше — ckanext-s3filestore и другие	ядро	ядро (DataStore); заливка CSV — расширение XLoader	перезаписывается; новый способ работы с файлами в 2.12 — без изменения содержимого
Dataverse	ядро	ядро: S3 и совместимые, Swift, Globus; несколько хранилищ	ядро: доверенное удалённое хранилище	ядро: разбор табличных файлов, описание переменных	замена создаёт новую версию набора
InvenioRDM, Zenodo	ядро	ядро с v13: S3 и совместимые	ядро: удалённый файл, без гарантий	нет	Zenodo — правка 30 дней, дальше новая версия
DSpace	ядро	ядро: S3; несколько хранилищ	нет	нет	хранится контрольная сумма
Huwise (OpenDataSoft)	ядро: вложения к набору	хранит поставщик	не проверено	ядро: набор и есть таблица записей	не проверено
Socrata	ядро	хранит поставщик	ядро: запись «ссылка на внешний источник»	ядро	не проверено
ArcGIS Hub	ядро	хранит поставщик	ядро: слой может жить на сервере ArcGIS	ядро (сервисы ArcGIS)	не проверено
GeoNetwork 4.4	ядро: вложения к записи в каталоге данных сервера	модули S3, JCloud и CMIS в поставке	ядро, основной режим	нет	не проверено
udata (data.gouv.fr)	ядро	не проверено	ядро	отдельные открытые сервисы hydra и api-tabular	не проверено
PxWeb, .Stat Suite	нет	—	нет	ядро (у PxWeb — собственный API, у .Stat — SDMX)	—
Статический сайт, Git	файлы рядом со страницами или по ссылке	как решите	как решите	нет, пока сами не сделаете	история Git хранит все версии

Для облачных сервисов «хранит поставщик» означает, что владелец портала не выбирает, где лежат байты, и не может забрать их иначе как выгрузкой. Это нужно учитывать в договоре: формат и срок выгрузки при уходе — такое же требование, как цена. Подробнее — в главе «Свои силы, свободное ПО или облачный сервис».

Расширения CKAN и их риски

У CKAN хранение — та область, где расширений больше всего и они устаревают быстрее всего. Методика оценки расширения — в главе об экосистемах расширений, здесь то, что нашлось при сверке.

- **Облачное хранилище.** В CKAN 2.12 появились настраиваемые хранилища, но в поставке только хранение на диске. Адаптеры облаков вынесены в ckanext-file-keeper-cloud сообщества CKAN; его тесты идут только на 2.12, выпусков на GitHub пока нет. Для версий до 2.12 есть расширения прежнего поколения: ckanext-s3filestore компании Keitaro тестируется на 2.10 и 2.11, но не на 2.12; ckanext-cloudstorage не меняли с февраля 2024 года и автоматических тестов у него нет; ckanext-blob-storage компании Datopian тестируется на 2.8 и 2.9, последний выпуск — январь 2023 года. Переход на 2.12 для портала с файлами в облаке — это смена расширения хранения, а не просто обновление.
- **Публичный бакет.** Документация CKAN 2.12 предупреждает: если файлы ресурсов и картинки организаций лежат в одном хранилище и оно открыто на чтение, файлы закрытых наборов доступны всем без проверки прав. Для ресурсов и публичных картинок заводят разные хранилища.
- **Загрузка таблиц в DataStore.** Официально поддерживаемый загрузчик — XLoader, тесты на 2.10–2.12. Альтернатива с разбором типов и вычислением полей описания — DataPusher+ компании datHere, его тесты идут на CKAN 2.11, выпуск 3.0.0 — январь 2026 года. Прежний DataPusher больше не поддерживается, документация CKAN предлагает уйти с него.
- **Проверка ссылок и копии.** ckanext-archiver и работающий поверх него ckanext-qa, который оценивает открытость формата, ведёт сообщество CKAN, тесты на 2.10–2.12.

На живых порталах эти сборки видны по адресу `/api/3/action/status_show`. 12 октября 2026 года data.gov.gr на CKAN 2.11.3 держал файлы в Azure через cloudstorage и свой модуль, проверял ссылки archiver и qa, таблицы загружал XLoader. opendata.swiss на 2.11.6 и HDX на 2.11.4 хранили файлы в S3 через s3filestore, таблицы загружали собственная версия XLoader и DataPusher+. Ирландский data.gov.ie работал на CKAN 2.9.16 с прежним DataPusher — версия ядра и загрузчик, которые уже не поддерживаются.

Публикация без портала

Без портала хранилище выбирают сами, и вариантов больше, чем кажется.

- **Файлы рядом со страницами** статического сайта — для небольших наборов: всё лежит в одном месте и переезжает вместе с сайтом.
- **Git-репозиторий** — история всех версий бесплатно, но предел 100 МиБ на файл. Большие файлы — во вложения к выпуску или в Git LFS. Подробнее — в главе о Git-репозиториях.

- **Zenodo или другой универсальный репозиторий** — когда нужна неизменная копия с DOI. Выпуск репозитория GitHub можно архивировать в Zenodo автоматически.
- **Открытый бакет объектного хранилища** — для больших данных; рядом нужен файл `README.md` с описанием и лицензией. Подробнее — в главе об облачных хранилищах и хабах.

Правило то же, что у портала: адрес файла не должен меняться, а исправленный файл — получать новое имя или новую версию. Контрольные суммы можно положить рядом файлом `SHA256SUMS` или записать в `datapackage.json`.

Как проверить живой портал

Ручная проверка занимает десять минут.

1. Откройте пять-десять наборов разных организаций и нажмите «Скачать». Посмотрите адрес файла.
2. Домен совпадает с порталом или это облачное хранилище портала — файл лежит у каталога. Домен другой — это ссылка на источник.
3. Откройте несколько внешних ссылок. Посмотрите, что открылось: файл, страница со списком файлов, главная страница сайта или ошибка.
4. Если на странице набора есть таблица с фильтром и отдельный адрес API для её строк — включено хранилище таблиц.
5. Найдите контрольную сумму или дату последней проверки ссылки. Если их нет, каталог не знает, тот ли это файл и жив ли он.

Замер на четырёх порталах

На большом портале то же делают программой. 12 октября 2026 года через API CKAN взята тысяча наборов на каждом из четырёх порталов: десять блоков по сто наборов, равномерно по всему каталогу. Затем из каждого набора взят один случайный внешний ресурс, всего 300 на портал, и каждую ссылку открыли. Неудачные запросы повторили по одному с ожиданием до минуты.

Портал	Наборов	Ресурсов в выборке загружено на портал	Наборов с таблицей в DataStore	Внешние ссылки: открылись	Ошибка или ресурса нет	Закрыто для программ	Страница вместо файла
data.gov.gr (Греция)	22 736	35%	3%	94%	5%	1%	1%
HDX (ООН)	28 312	55%	7%	68%	6%	24%	2%
data.gov.ie (Ирландия)	22 770	0%	41%	95%	3%	меньше 1%	1%
opendata.swiss (Швейцария)	16 521	0%	2%	94%	2%	0%	4%

«Ошибка или ресурса нет» объединяет ответы 404 и 410, ошибки сервера, недоступные домены, таймауты и ошибки сертификата. «Закрыто для программ» — ответы 401, 403 и 429: файл может существовать, но программа его не получает. «Страница вместо файла» — у

ресурса указан формат файла (CSV, XLSX, PDF), а открывается веб-страница. Выборка в 300 ссылок даёт точность в несколько процентных пунктов, это ориентир, а не перепись. Все загруженные файлы, которые попали в контрольную выборку (по сто на греческом портале и HDX), открылись.

Что из замера следует.

Ошибки сосредоточены у немногих источников. На греческом портале 12 из 300 ссылок не открылись из-за ошибки сертификата, и все они ведут на один картографический сервер муниципалитета Ираклиона. На HDX 73 ссылки из 300 ответили отказом 403, и все — на каталог микроданных УВКБ ООН, который закрыт от программ. Проверку ссылок поэтому полезно сводить по доменам: один разговор с одним источником убирает большую часть проблем.

«Закрыто» и «удалено» — разные проблемы. Файл за проверкой от роботов существует, человек его скачает. Но сборщик, веб-архив и программа пользователя его не получат, и для них это такая же пропажа. Канадский open.canada.ca в тот же день отвечал на автоматические запросы отказом и задержками и на свой собственный служебный адрес API, поэтому в таблицу он не вошёл. Подробнее — в антипаттерне «Дверь закрыта для программ».

Ссылка открылась — ещё не значит, что данные на месте. Восемь наборов из выборки на швейцарском портале указывают формат CSV, а ссылка у всех ведёт на одну и ту же страницу города Берна со списком файлов. Ссылка с форматом PDF на сайт statistiquedesrevenus.ch переадресует на раздел сайта Agroscope: старый домен перенаправили, и вместо документа открывается раздел сайта. Программа, которая проверяет только код ответа, посчитает обе ссылки рабочими.

Хорошо собранный портал ссылок может быть надёжным. У ирландского и швейцарского порталов ни одного своего файла в выборке, а явных пропаж меньше 3%. Основные источники — статистические службы и городские порталы — сами ведут свои данные и не меняют адреса. Цифра data.gov.uk о четверти нерабочих ссылок — результат многолетнего сбора ссылок на сайты сотен муниципалитетов без проверки, а не неизбежное свойство модели ссылок.

Команды

Для портала на CKAN. Первая команда берёт 500 наборов и считает долю загруженных ресурсов, ресурсов с таблицей в DataStore и самые частые внешние домены:

```
PORTAL=https://data.gov.gr
curl -s "$PORTAL/api/3/action/package_search?rows=500&start=5000" \
  | jq '[.result.results[].resources[]] as $r | ($r | length) as $n | {
      resources: $n,
      uploaded: ([$r[] | select(.url_type == "upload")] | length / $n),
      in_datastore: ([$r[] | select(.datastore_active == true)] | length / $n),
      top_external_hosts: ([$r[] | select(.url_type != "upload") | .url |
capture("^https?://(?<h>[^/]+)").h]
        | group_by(.) | map({host: .[0], n: length}) | sort_by(-.n) | .[0:5])
    }'
```

Вторая открывает внешние ссылки из той же выборки и печатает код ответа, тип содержимого и конечный адрес. Код 200 с типом `text/html` у ресурса с форматом CSV — повод открыть ссылку глазами:

```
curl -s "$PORTAL/api/3/action/package_search?rows=50&start=5000" \
  | jq -r '.result.results[].resources[] | select(.url_type != "upload") | .url' \
  | head -30 \
  | while read -r url; do
      curl -s -o /dev/null -L -m 30 -A "Mozilla/5.0" \
        -w "%{http_code}\t%{content_type}\t%{url_effective}\n" "$url"
    done
```

Не все серверы отвечают на запрос `HEAD`, поэтому команда делает обычный запрос и не сохраняет тело ответа. Запускайте проверку медленно, по одной ссылке: десятки параллельных запросов к одному сайту выглядят как атака, и вы получите отказ, которого у обычного пользователя не было бы.

Чтобы узнать, какие расширения хранения и проверки ссылок стоят на портале:

```
curl -s "$PORTAL/api/3/action/status_show" \
  | jq '{ckan_version: .result.ckan_version,
         storage: [.result.extensions[] |
select(test("xloader|datapusher|archiver|qa|cloud|s3|blob|filestore|file_keeper"))]}'
```

Для data.gouv.fr результаты проверки уже записаны в описании каждого ресурса: поля `check:status`, `check:available`, `check:date` и контрольная сумма `analysis:checksum`. Доля загруженных на портал файлов видна по полю `filetype`: `file` — у портала, `remote` — ссылка.

```
curl -s "https://www.data.gouv.fr/api/1/datasets/?page_size=100&page=200" \
  | jq '[.data[].resources[]] as $r | ($r | length) as $n | {
      resources: $n,
      uploaded: ([$r[] | select(.filetype == "file")] | length / $n),
      checked: ([$r[] | select(.extras["check:status"] != null)] | length / $n),
      not_available: ([$r[] | select(.extras["check:available"] == false)] | length / $n)
    }'
```

Как превратить доли нерабочих ссылок в метрику качества и отслеживать её — в главе о метриках качества.

Что из этого следует

- Решите до запуска, где лежат байты: диск, облако или ссылки. Перенос уже загруженных файлов дороже, чем выбор в первый день.
- Если на ваши файлы ссылаются годами — храните их у себя, не перезаписывайте после публикации и считайте контрольные суммы. Исправление — новая версия.
- Если каталог ссылается наружу — проверяйте ссылки по расписанию, сводите ошибки по источникам и храните копии важных наборов. Проверяйте не только код ответа, но и то, что открылся файл, а не страница.
- Поднимите предел размера загрузки до реального: 10 МиБ по умолчанию в CKAN незаметно превращают портал в портал ссылок.

- Таблицу в базе с API заводите для таблиц, которые фильтруют и запрашивают, и всегда оставляйте рядом выгрузку целиком.
- В CKAN выбирайте расширения хранения под свою версию ядра: при переходе на 2.12 облачное хранилище меняет расширение.
- Давайте пользователю постоянный адрес файла на портале, а не временную ссылку облачного хранилища.

Источники

- CKAN 2.12: хранение файлов и настраиваемые хранилища — предел 10 МиБ, адаптеры облаков в отдельном расширении, предупреждение о публичном хранилище, неизменяемые файлы нового способа; DataStore — PostgreSQL, пределы выдачи, выгрузка, XLoader, DataPusher+ и прекращение поддержки DataPusher
- Расширения CKAN: ckanext-file-keeper-cloud, ckanext-s3filestore, ckanext-cloudstorage, ckanext-blob-storage, XLoader, DataPusher+, ckanext-archiver, ckanext-qa — файлы автоматических тестов и даты изменений, проверено 12 октября 2026 года
- Dataverse: хранение файлов — S3, Swift, Globus, доверенное удалённое хранилище, прямая загрузка, временные ссылки, контрольные суммы
- InvenioRDM: хранение в S3 (с v13) и способы передачи файлов — свой, забрать по ссылке, частями, удалённый
- Zenodo: работа с файлами — 50 ГБ и 100 файлов, правка в течение 30 дней
- DSpace: слой хранения — диск и S3, несколько хранилищ, контрольные суммы, плата за трафик
- GeoNetwork: модули хранения в выпуске 4.4.12
- Socrata: ссылка на внешний источник данных
- data.gouv.fr: обходчик hydra и табличный API; поля проверки в API наборов, проверено 12 октября 2026 года
- Замер 12 октября 2026 года: API CKAN порталов data.gov.gr, HDX, data.gov.ie, opendata.swiss, по 1000 наборов и 300 внешних ссылок на портал; адрес `status_show` для версий и расширений; переадресация HDX на временную ссылку S3
- И. Бегтин, Хорошие и плохие практики публикации данных, ноябрь 2024 — доля нерабочих ссылок на data.gov.uk при выгрузке 2023 года
- Congressional Research Service, Data.gov: Implementation and Perspectives on Its Functions, R48954, 21 мая 2026 года — реестр или хранилище, сохранность доступа к данным
- GDS, What's changing on data.gov.uk and why, 25 марта 2026 года — не меньше 25% ссылок ведут на ошибку
- Анатомия каталога, API и массовая выгрузка, публикация без портала: статические сайты — data.gov.cz как портал без файлов
- Передовые каталоги — табличный API data.gouv.fr и перевод наборов в Parquet на Hugging Face

Постоянные идентификаторы, DOI и версии

На данные ссылаются: в статье, в отчёте, в другом каталоге, в чужой программе. Ссылка полезна, только если через год она открывает тот же набор. Отсюда три практических вывода главы.

Постоянный адрес страницы нужен каждому каталогу и почти ничего не стоит, если заложить его до запуска: в адресе должен стоять ключ, который не меняется при правке описания и при переезде на другую платформу. DOI нужен там, где на данные ссылаются как на источник. Это не кнопка, а договор с агентством регистрации и обязательство годами поддерживать страницу, на которую он ведёт. Версия — обещание, что ссылка откроет те данные, которыми пользовался автор, а не сегодняшнюю выгрузку.

Главное, что показывают проверки в конце главы: идентификатор не надёжнее сайта, на который он указывает. В DataCite больше пятисот «находимых» DOI наборов данных ведут на `localhost` — адрес компьютера, на котором их настраивали. У четырёх из десяти проверенных установок DSpace постоянный идентификатор, напечатанный на странице записи, не открывается.

Что это

Адрес и идентификатор. Адрес говорит, где лежит страница: `portal.example/dataset/budget-2024`. Идентификатор говорит, что это за объект, и не зависит от того, где объект сейчас. Постоянный идентификатор состоит из трёх частей: имени, службы, которая по имени отправляет на текущий адрес (её называют резолвером), и владельца, который обязуется обновлять этот адрес при каждом переезде. Без третьей части первые две не работают: служба честно перенаправит туда, куда ей сказали, даже если там давно ничего нет.

Механизмов три, и их часто путают.

Постоянный адрес страницы. Адрес набора не меняется из-за правки опечатки в описании, не содержит сегодняшнюю дату и номер строки в базе. При смене сайта старый адрес перенаправляет на новый. Это нужно всем, и обеспечивает это сам каталог.

Внешний постоянный идентификатор. DOI, Handle, ARK и подобные: имя регистрируют во внешней службе, а она перенаправляет на текущую страницу набора. Ссылка вида `https://doi.org/10.xxxx/yyyy` переживает смену домена, если владелец не забыл обновить адрес в службе.

Версия. Зафиксированное состояние набора. Научная ссылка указывает на ту версию, которую использовали в расчёте. Портал открытых данных чаще ведёт один обновляемый набор: важны дата обновления и архив прошлых выгрузок, а не новый DOI на каждый понедельник.

Какие бывают идентификаторы

Идентификатор	Кто выдаёт и сколько стоит	Что обязательно	Где встречается
DOI	агентства регистрации: DataCite для данных, Crossref для публикаций и другие. Платит организация-участник, обычно через консорциум	описание по схеме агентства: у DataCite шесть обязательных полей	научные репозитории, журналы, Zenodo, Figshare, некоторые геокаталоги
Handle	Handle.Net Registry; префикс — за ежегодный сбор, свой сервер Handle	нет	DSpace по умолчанию, Dataverse как вариант. Технически DOI — тоже Handle с префиксом `10.`
ARK	ARK Alliance; номер организации (NAAN) бесплатно для библиотек, архивов и подобных «организаций памяти»	нет	национальные библиотеки и архивы, например Национальная библиотека Франции
PURL, w3id.org	бесплатные службы перенаправления: purl.org ведёт Интернет-архив, w3id.org — сообщество W3C, правила добавляют запросом в GitHub	нет	словари и онтологии связанных данных, постоянные адреса небольших проектов
CSTR	национальная система Китая	описание объекта	китайские научные репозитории, часто вместе с DOI
Постоянная ссылка установки	сама платформа, без внешней службы	нет	Dataverse (PermaLinks), идентификатор записи в любом каталоге

Рядом с идентификатором набора работают идентификаторы людей и вещей вокруг него: ORCID у автора, ROR у организации, IGSN у физического образца, RAiD у исследовательского проекта. IGSN и RAiD с 2020-х годов выдаются как DOI через DataCite: среди её репозиториев 40 каталогов образцов и 17 реестров проектов. По этим связям агрегаторы собирают граф: кто сделал набор, в какой организации, по какому проекту и в какой статье его использовали.

Зачем нужна

Ссылки умирают быстро. По исследованию Pew Research Center 2024 года, из страниц, которые существовали в 2013 году, к октябрю 2023-го не открывались 38%. На 21% страниц государственных сайтов США есть хотя бы одна битая ссылка, чаще всего на сайтах городов. С наборами данных так же: при полной выгрузке data.gov.uk в 2023 году не открывались около 10% ссылок на файлы, подробнее — в главе о хранилище.

Ссылки ломают люди, а не технологии. Так в заметке «Cool URIs don't change» (1998) писал Тим Бернерс-Ли, и W3C с 1999 года держит все свои документы по прежним адресам по собственной политике постоянства. Обратный пример — российский реестр отечественного ПО в 2021 году: после переезда старые ссылки перестали открываться, а в новых адресах стоял технический номер записи в базе (330494) вместо номера программы в реестре (10269). При следующей смене базы такой адрес сломается снова. Короткие ссылки тоже не постоянны: их судьбу решает владелец сервиса, как показала история goo.gl в главе об архивации.

Чтобы на данные сослались и это засчитали. DOI делает набор полноправным источником: его ставят в список литературы, по нему считают цитирования, по нему агрегаторы связывают набор со статьёй и автором. Без идентификатора ссылка на данные остаётся адресом сайта, который никто не считает. Как из этого получается готовая строка ссылки — в главе об экспорте цитирования.

Чтобы сопоставлять записи. Реестры, справочники и классификаторы нужны другим данным как точка отсчёта. Если у записи нет устойчивого идентификатора, нельзя надёжно понять, что строка в прошлогоднем файле и строка в новом — одна и та же организация. Об этом раздел о реестрах ниже.

Как устроен DOI

DOI состоит из префикса и суффикса: `10.5281/zenodo.16546069`. Префикс `10.5281` выдан организации, суффикс она придумывает сама. Служба doi.org по запросу отвечает перенаправлением на адрес, который владелец записал при регистрации. Этот адрес называют посадочной страницей: на ней описание набора, ссылка для цитирования и путь к файлам. Сам DOI файлов не хранит.

Агентства. Для данных DOI регистрирует DataCite — объединение, зарегистрированное в Германии. 12 октября 2026 года в DataCite 138,3 миллиона DOI, из них 76,4 миллиона с типом «набор данных». Что стоит за этим числом, разобрано в главе о мире каталогов в цифрах. Crossref, агентство из США, регистрирует DOI статей, книг и препринтов. Есть и другие агентства, например Бюро публикаций ЕС. Начиная с InvenioRDM v14 и в Dataverse можно регистрировать DOI и через Crossref, но для наборов данных обычный выбор — DataCite.

Описание обязательно. При регистрации в DataCite заполняют шесть полей: идентификатор, автор, название, издатель, год, тип ресурса. Без автора и года из DOI не получится ссылка для цитирования. Что ещё стоит заполнить, сказано в главе о полноте метаданных.

DOI нельзя удалить. У записи в DataCite три состояния. Черновик ещё не зарегистрирован в глобальной системе, его можно удалить. «Зарегистрированный» DOI работает, но его описание скрыто из поиска. «Находимый» работает и виден всем. Из двух последних состояний пути назад нет. Если набор пришлось снять, DataCite советует оставить на его месте страницу-надгробие: ссылка для цитирования, DOI и объяснение, почему данных больше нет. Для тех, кто не может сделать такую страницу, у DataCite есть общая страница-заглушка; 12 октября 2026 года на неё вели 5145 находимых DOI.

Сколько стоит. DataCite не берёт плату за каждый DOI, пока организация регистрирует меньше 100 тысяч в год. Платят за участие: самостоятельный член — 2000 евро в год плюс инфраструктурный сбор от 200 до 700 евро в зависимости от страны. Организация внутри консорциума членский взнос не платит, только инфраструктурный сбор. Поэтому университеты обычно получают DOI через национальный или библиотечный консорциум, а небольшие проекты — через чужой репозиторий вроде Zenodo, где DOI бесплатен для автора.

Чего DOI не гарантирует

Что данные сохранятся. DOI ведёт на страницу, а страницу держит владелец. Исследование «Open is not forever» (2020) насчитало 176 журналов открытого доступа, которые исчезли с 2000 по 2019 год, — с DOI и без. Идентификатор помогает найти след, но не заменяет архив. Как сохранять сами данные, разобрано в главе об архивации.

Что адрес правильный. Служба doi.org перенаправляет туда, куда записал владелец, и не проверяет, есть ли там что-нибудь. 12 октября 2026 года в DataCite 765 находимых DOI ведут на `localhost`, из них 548 — наборы данных. Такой адрес открывается только на компьютере того, кто его записал. Одна из причин видна в документации GeoNetwork: в качестве шаблона адреса посадочной страницы она предлагает `http://localhost:8080/geonetwork/srv/resources/records/{{uuid}}`. Португальский институт моря и атмосферы IPMA в 2025 году зарегистрировал два DOI с таким адресом, один из них — с названием «Copy of template MFSD Template…», то есть тестовую копию шаблона.

Что за DOI стоит набор. DOI выдают и отдельной записи большой базы: наблюдению вида, кристаллической структуре, отдельному эксперименту. Поэтому число DOI с типом «набор данных» нельзя читать как число наборов, см. главу о мире каталогов в цифрах.

Что агентство будет работать с вами всегда. DataCite и Crossref соблюдают санкции своих стран. Crossref не принимает заявки от организаций, которые входят в правительство России или находятся под санкциями США, Великобритании или ЕС; остальные заявки из России и Беларуси проходят отдельную проверку, а после 2022 года каждого российского члена проверили по санкционным спискам и часть исключили. Уже выданные DOI продолжают разрешаться, но новые организация может не получить. Китай выстроил параллельную систему: национальный идентификатор CSTR и репозиторий Science Data Bank, который при этом выдаёт DOI через DataCite (28 406 записей на октябрь 2026 года) и отдаёт описания европейским и американским агрегаторам.

Версии

Что считать новой версией. Изменились данные — новая версия. Поправили опечатку в описании или добавили ключевое слово — ревизия, новой версии она не создаёт. Граница проходит по тому, получит ли читатель другой результат, если повторит расчёт.

В научном репозитории опубликованная версия не переписывается. Исправление — новая версия со своим DOI или с тем же DOI и новым номером, по правилам репозитория. Платформы делают это по-разному, и разница важна для ссылки.

Модель	Как устроено	Где
DOI на каждую версию и общий DOI	у каждой версии свой DOI, плюс общий (родительский) DOI, который всегда ведёт на последнюю версию	InvenioRDM и Zenodo, Figshare (общий DOI и DOI с суффиксом `.v1`, `.v2`)
Один DOI и номер версии	DOI один на весь набор, версия (1.0, 1.1, 2.0) указывается в ссылке для цитирования	Dataverse
DOI с номером версии	DOI новой версии — исходный DOI с номером через точку	DSpace при включённых версиях записей

Модель	Как устроено	Где
Один обновляемый набор и снимки на дату	адрес набора постоянный, файл заменяется; прошлые выгрузки лежат в архиве с датой	порталы открытых данных, реестры, справочники
Серия наборов	каждый выпуск — отдельный набор со своим адресом, серия объединяет их	класс `dcat:DatasetSeries` в DCAT 3, статистика по годам

У InvenioRDM и Figshare статья может сослаться на DOI конкретной версии, а общий DOI остаётся тем, кому нужна всегда последняя. У Dataverse статья ссылается на DOI и номер версии вместе. InvenioRDM с версии 13 позволяет сделать DOI необязательным, но по умолчанию не даёт выпустить версию без DOI, если у предыдущей он был: иначе общий DOI мог бы привести к версии без идентификатора.

Одна ссылка — разное содержимое. На портале открытых данных ссылка на файл обычно ведёт на текущий файл. В CKAN адрес загруженного файла строится по идентификатору ресурса, имя файла в нём ни на что не влияет: заменили файл — по старой ссылке скачивается новый. Для портала это удобно, для научной ссылки — нет: через год по ней получат не те данные. Помогают две вещи. Контрольная сумма показывает, что файл не изменился: в DCAT 3 для неё есть свойство у дистрибутива, Dataverse дополнительно считает для таблиц UNF — контрольную сумму содержимого, не зависящую от формата файла. И снимки на дату: если на файл уже могли сослаться, прошлый вариант остаётся в архиве.

Как версию описывают метаданные. В DCAT 3 (2024) появились свойства версии: номер (`dcat:version`), что изменилось (`adms:versionNotes`), предыдущая и текущая версии (`dcat:previousVersion`, `dcat:hasCurrentVersion`). В DataCite версии связывают отношениями `IsNewVersionOf`, `IsVersionOf`, `HasVersion`. У CKAN есть поле «версия», и ckanext-dcat выгружает его в DCAT, но это просто подпись: CKAN не хранит прошлые состояния набора как отдельные объекты со своими адресами.

На портале открытых данных набор обычно один, а файлы заменяются. Тогда нужны:

- дата и периодичность обновления в описании;
- неизменный адрес набора при замене файла;
- архив предыдущих файлов с датой, если на них уже могли сослаться.

Реестры и справочники

Реестр лицензий, перечень организаций, классификатор видов деятельности — наборы, на которые ссылаются другие данные и программы. Для них постоянный адрес нужен не только набору, но и каждой записи. Их ведут одним из двух способов.

- **Справочник с версиями.** Публикуется целиком, каждая версия — отдельный файл с датой. Так ведут классификаторы и адресные базы: пользователь берёт версию на нужную дату и может сравнить две версии между собой.
- **Учётная система с историей записи.** Записи меняются по одной, у каждой хранится история правок. Так удобнее для реестров, где изменения идут каждый день: лицензии, разрешения, статусы.

В обоих случаях у реестра, которым удобно пользоваться, есть одно и то же:

- уникальный идентификатор у каждой записи, присвоенный по правилам реестра, а не номер строки в файле и не технический номер в базе;
- постоянная ссылка на реестр целиком и на каждую запись, построенная из этого идентификатора;
- записи не удаляются, а получают статус «исключена» с датой;
- видны прошлые версии записи или всего реестра;
- указано основание записи: решение, приказ, документ, со ссылкой;
- реестр выгружается целиком, регулярными снимками на дату, и доступен через API.

Опубликованный реестр без идентификаторов и истории — снимок, который нельзя надёжно сопоставить с другими данными: запись могли переименовать, исключить или внести с ошибкой, и по файлу этого не видно. Регулярные снимки к тому же дешевле для самой системы: выгрузка всего реестра по запросу на большой базе легко кладёт сервер, а готовый файл раз в сутки — нет.

Идентификаторы объектов внутри данных

Тот же подход работает и для данных, которые реестром не называют. У Overture Maps постоянный идентификатор GERS есть у каждого здания, дороги, места и административной единицы в ежемесячном выпуске. Рядом ведутся реестр идентификаторов с датами первого и последнего появления, журнал изменений по объекту и таблицы соответствия с идентификаторами источников. Другой поставщик может публиковать только свои показатели — пробки, население — без геометрии, а соединяют их по идентификатору.

Урок там же. В выпуске 2025-06-25.0 Overture заменил прежние идентификаторы, в которых были зашиты ячейка сетки H3 и тема, на случайные UUID, и старые номера больше не находятся. Идентификатор со смыслом внутри ломается, когда меняется смысл. У OpenSanctions после устранения дублей у сущности один канонический идентификатор, а если она совпала с записью Викиданных — номер Викиданных. Подробнее — в главе о передовых каталогах.

Варианты реализации

Постоянный адрес силами платформы

Главный вопрос — какой ключ стоит в адресе набора. Короткое имя из названия удобно читать, но его хочется поправить вместе с названием. Номер или код, присвоенный один раз, читать неудобно, зато он не меняется.

Платформа	Что в адресе	Что будет при переименовании
CKAN	короткое имя набора (`/dataset/budget-2024`); у набора есть и неизменный UUID	адрес по старому имени отвечает «не найдено». Адрес по UUID всегда перенаправляет на текущее имя — его и стоит давать для ссылок
Socrata	код из двух групп по четыре символа вида `erm2-nwe9` и название	название в адресе декоративное: любой путь с кодом перенаправляет на текущий адрес
Huwise (OpenDataSoft)	читаемый идентификатор набора; есть и внутренний, который не меняется никогда	читаемый идентификатор можно менять только до публикации

Платформа	Что в адресе	Что будет при переименовании
Dataverse	постоянный идентификатор (DOI, Handle или постоянная ссылка установки)	не зависит от названия
DSpace	Handle	не зависит от названия
GeoNetwork	UUID записи	не зависит от названия
InvenioRDM	идентификатор записи	не зависит от названия; у каждой версии свой

Проверено 12 октября 2026 года: на data.gov.gr адрес набора по UUID отвечает перенаправлением на адрес по имени, а на портале Нью-Йорка на Socrata путь с произвольным названием и кодом набора приводит к текущей странице.

Свой постоянный адрес без DOI

Если DOI не нужен, а адрес должен пережить смену платформы, есть недорогие пути.

- **Перенаправления на своём домене.** Адреса наборов строятся от домена, который организация не собирается менять, и при каждом переезде на веб-сервере ставят таблицу перенаправлений «старый адрес → новый». Самый дешёвый способ и единственный, который нужен всем.
- **Handle.** Префикс получают в Handle.Net Registry за ежегодный сбор и держат свой сервер Handle. DSpace поставляется с ним, но с префиксом-заглушкой `123456789`, который глобальная система не знает. Пока префикс не зарегистрирован и сервер не запущен, Handle на страницах — только видимость.
- **ARK.** Номер организации выдаётся бесплатно, ARK можно разрешать на своём сервере или через общий резолвер n2t.net. Подходит библиотекам, архивам, музеям.
- **w3id.org и PURL.** Бесплатные службы перенаправления. На w3id.org правило для своего пространства имён добавляют запросом в репозиторий на GitHub. Подходят небольшим проектам и статическим сайтам, которым нужен адрес, не привязанный к хостингу.

DOI

В ядре платформы. Dataverse, InvenioRDM, DSpace, Figshare и GeoNetwork регистрируют DOI сами. Нужны договор с агентством, префикс и тестовая учётная запись. InvenioRDM и DSpace по умолчанию подключены к тестовой системе DataCite, чтобы не выпустить настоящие DOI по ошибке. Ошибку, допущенную уже в рабочей системе, исправить трудно: DOI не удаляются.

Через расширение. У CKAN DOI нет в ядре. Расширение ckanext-doi Музея естественной истории в Лондоне переводит поля набора в схему DataCite и регистрирует DOI. Про риски — в разделе о расширениях.

Вручную или через консорциум. Организация с префиксом может регистрировать DOI в веб-интерфейсе DataCite Fabrica или через его API и указывать в них адреса страниц любой платформы, в том числе облачной. Так делают, когда DOI нужен десятку наборов, а не каждому. Консорциумы предлагают это как услугу: например, австралийская ARDC выдаёт DOI для данных государственных научных организаций, но прямо предупреждает, что поддерживать адреса придётся самим.

В чужом репозитории. Копия или выпуск данных уходит в Zenodo или Figshare и получает DOI там. Так делают с репозиториями на GitHub: выпуск архивируется в Zenodo автоматически. Подробнее — в главе о Git-репозиториях. Минус один: DOI ведёт на копию в Zenodo, а не на ваш портал, и обновлять копию нужно отдельно.

Когда нужна, а когда нет

Ситуация	Что нужно	Чем обычно закрывают
Научный репозиторий, грантодатель или журнал требует ссылку на данные	DOI на каждую опубликованную версию, ORCID авторов, страница-надгробие при снятии	Dataverse, InvenioRDM, DSpace с DataCite, Figshare, Zenodo
Портал открытых данных министерства или города	постоянный адрес по неизменному ключу, перенаправления при переезде, дата обновления; DOI — не нужен	любая платформа, если выбрать ключ в адресе
Расписание транспорта, оперативные данные, которые меняются ежедневно	постоянный адрес набора и ленты; DOI лишний	любая платформа
Реестр или справочник	идентификатор каждой записи, снимки на дату, история	своя учётная система и выгрузки; портал — для снимков
Геопортал с записями ISO	UUID записи в адресе; DOI — если на слои ссылаются в научных работах	GeoNetwork с включённым DOI
Отдельные наборы, на которые ссылаются в докладах организации	DOI только этим наборам	Fabrica по договору консорциума или копия в Zenodo
Публикация без портала	адрес, не привязанный к хостингу, DOI на выпуск	Git-репозиторий и Zenodo, w3id.org

DOI стоит выдавать, когда на набор будут ссылаться как на источник и организация готова годами держать посадочные страницы. DOI не стоит выдавать «для солидности»: ежедневно меняющийся набор с DOI вводит в заблуждение — ссылка обещает зафиксированное состояние, а открывает сегодняшнее.

Коротко, как выглядит обоснование по шаблону карточки решения:

- **Контекст.** Университет ведёт портал на CKAN 2.12 для данных кафедр. Фонд, который финансирует большую часть проектов, начал требовать DOI для данных к отчётам.
- **Варианты.** Поставить ckanext-doi; завести рядом Dataverse или InvenioRDM для данных, на которые ссылаются; отправлять такие наборы в Zenodo; регистрировать DOI вручную через консорциум.
- **Критерии.** Неизменность опубликованной версии, поддержка расширения на текущей версии CKAN, стоимость сопровождения, где окажется ссылка — на своём портале или в чужом репозитории.
- **Решение.** Рядом с CKAN поставить Dataverse для данных к публикациям; CKAN оставить для оперативных и учебных наборов; DOI получать через библиотечный консорциум.
- **Почему.** CKAN не хранит версии: DOI на заменяемый файл не даёт той неизменности, ради которой фонд его требует. ckanext-doi на CKAN 2.11 и 2.12 не тестировался. Zenodo проще, но ссылки уводят с портала университета, а копии приходится обновлять вручную.

- **Когда пересмотреть.** Если ckanext-doi начнут тестировать на текущей версии CKAN и потребуется DOI только для небольшого числа наборов без версий, либо если данных к публикациям станет так мало, что хватит Zenodo.

Как это сделано в платформах

Пометки — ядро, расширение, нет, не проверено, как в сводной таблице. Сверено с документацией платформ 12 октября 2026 года.

Платформа	DOI	Версии	Что важно знать
Dataverse 6.12.1	ядро: DataCite, Crossref (с 6.12.1), EZID, Handle и постоянные ссылки установки; несколько поставщиков на одной установке	ядро: основные и малые версии (1.0, 1.1, 2.0), DOI один на набор	файлам можно выдавать собственные идентификаторы — зависимые от DOI набора или независимые
InvenioRDM v14, Zenodo	ядро: DataCite, с v14 — Crossref; можно принять DOI, выданный в другом месте	ядро: у каждой версии свой DOI плюс родительский (с v12)	с v12 не регистрирует DOI для записей с ограниченным доступом; с v13 DOI можно сделать необязательным
Figshare	ядро сервиса: DataCite	ядро: общий DOI ведёт на последнюю версию, у версий суффикс `.v1`, `.v2`	облачный сервис, своя установка невозможна
DSpace 10	ядро, включается настройкой: DataCite или EZID, в дополнение к Handle	ядро: версии записей; DOI версии — исходный с номером через точку	адрес записи по-прежнему строится от Handle; DOI отправляются в DataCite отдельной задачей по расписанию; фильтр решает, каким записям DOI положен
GeoNetwork 4.4	ядро, включается в настройках: DataCite или API Бюро публикаций ЕС; только для публичной записи	нет: при включённом согласовании есть рабочая копия записи и история статусов, но не версии со своими идентификаторами	DOI получает описание слоя, а не данные: слой на картографическом сервере может меняться
CKAN 2.12	расширение ckanext-doi: заявлены CKAN 2.9 и 2.10	нет: журнал изменений и поле «версия» — не версии со своими адресами	для ссылок давайте адрес по UUID набора
NADA	нет: только поле для DOI, выпущенного отдельно	не проверено	цитирование через DataCite работает, если DOI есть
Huwise, Socrata, ArcGIS Hub	нет	нет	постоянный адрес — по неизменному коду набора или элемента; DOI можно выпустить вручную и указать в своём поле
Hugging Face, Kaggle	ядро площадки: DOI через DataCite по кнопке	Hugging Face — ревизии репозитория Git; DOI выдаётся на текущую ревизию, для новой выпускают новый, а старый помечается устаревшим	после выдачи DOI набор на Hugging Face нельзя удалить или переименовать без поддержки; подробнее — в главе об ML-каталогах
Статический сайт, Git-репозиторий	через Zenodo: DOI на каждый выпуск	метки выпусков в репозитории	ссылка на файл в ветке показывает текущее состояние, ссылка на метку выпуска — зафиксированное

Расширения и их риски

ckanext-doi — единственный путь к DOI внутри CKAN, и на октябрь 2026 года рискованный. Последний выпуск, 4.0.4, вышел 22 сентября 2025 года, с тех пор изменений нет. В описании заявлены CKAN 2.9 и 2.10. Автоматические тесты идут в готовых образах самого музея, а эти образы последний раз обновлялись 20 сентября 2023 года — до выхода CKAN 2.11. Значит, на 2.11 и 2.12 расширение не тестировалось вовсе. Портал, которому DOI нужен на текущем CKAN, либо берёт проверку на себя, либо выдаёт DOI вне CKAN. Состояние остальных расширений CKAN — в главе про CKAN.

Модули DOI в ядре тоже требуют внимания при обновлениях. В DSpace регистрация разделена на два шага: запись помечается в базе, а отправку в DataCite делает отдельная команда по расписанию. Если задачу по расписанию не перенесли на новый сервер при обновлении, DSpace считает DOI выданными, а в DataCite их нет, и ссылки не открываются. В GeoNetwork шаблон адреса посадочной страницы задают вручную, и с ним связана ошибка с `localhost`, описанная выше.

Что показывает реестр

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) пометка «DOI» стоит у 2008 каталогов из 43 614. Из них 1404 — научные репозитории, 584 — геопорталы, остальные единичны: 7 каталогов показателей, по 3 портала открытых данных и каталога микроданных.

Число нужно читать осторожно. Пометку в реестре в основном проставляют по платформе: если программа умеет выдавать DOI, пометку получают все её установки, у которых нет ручной отметки «нет». Поэтому среди геопорталов с DOI 555 — установки GeoNetwork. В самом DataCite 12 октября 2026 года 7209 DOI ведут на адреса, где в пути есть слово `geonetwork`, и приходятся они примерно на 45 установок; больше трёх тысяч — у французской гидрографической службы SHOM и австрийского ведомства геодезии BEV. Даже с поправкой на установки с другим адресом это на порядок меньше 555. Реестр показывает, кто может выдать DOI, а не кто выдаёт. Запросы — в приложении «Как работать с реестром».

Публикация без портала

Без портала постоянство адреса целиком на авторе, зато и способы проще.

- **Выпуск и DOI.** Репозиторий на GitHub связывают с Zenodo, и каждый выпуск получает свой DOI, а весь ряд выпусков — общий. Файл `CITATION.cff` в корне подсказывает, как ссылаться. Как это настроить, — в главе о Git-репозиториях.
- **Ссылка на версию, а не на ветку.** Ссылка на файл в ветке `main` всегда показывает текущее состояние. Для ссылки в статье берут адрес метки выпуска или DOI выпуска.
- **Адрес, не привязанный к хостингу.** Если набор живёт на статическом сайте, адрес вида `https://w3id.org/ваш-проект/` переживёт переезд с GitHub Pages на другой хостинг: поменяется только правило перенаправления.

Как проверить живой портал

- Измените описание набора на тестовом стенде, в том числе название. Адрес страницы не должен измениться, а если изменился, старый должен перенаправлять.
- Найдите на странице постоянную ссылку. Если её нет, посмотрите, есть ли в адресе неизменный ключ: UUID, код, номер.
- Если обещан DOI, он виден на странице, а `https://doi.org/…` открывает эту же страницу, а не главную и не `localhost`.
- Если на странице Handle, проверьте его через `hdl.handle.net`. Префикс `123456789` — заглушка из поставки DSpace.
- Откройте предыдущую версию набора и проверьте, что её номер стоит в предлагаемой ссылке.
- После миграции на старом адресе стоит перенаправление, а не страница «не найдено».

Команды

Куда ведёт DOI и что о нём знает DataCite:

```
curl -sI https://doi.org/10.5281/zenodo.16546069 | grep -i '^location'
curl -s https://api.datacite.org/dois/10.5281/zenodo.16546069 \
  | jq '.data.attributes | {url, state, publisher, version, related: [.relatedIdentifiers[]?
| select(.relationType | test("Version"))]}'
```

В первом ответе должно быть перенаправление на страницу набора. Во втором — адрес, который записал владелец, состояние и связи с другими версиями.

Сколько DOI у репозитория и есть ли среди них адреса, которые нигде, кроме компьютера владельца, не откроются. Идентификатор репозитория в DataCite (`client-id`) виден в ответе предыдущей команды, в разделе `relationships`:

```
CLIENT=cern.zenodo
curl -s "https://api.datacite.org/dois?client-id=$CLIENT&page%5Bsize%5D=0" | jq '.meta.total'
curl -s -G https://api.datacite.org/dois --data-urlencode "client-id=$CLIENT" \
  --data-urlencode 'query=url:*localhost*' --data-urlencode 'page[size]=0' | jq '.meta.total'
```

Разрешается ли Handle, напечатанный на странице записи DSpace:

```
curl -s https://hdl.handle.net/api/handles/1234/5678 | jq '.responseCode'
```

Код `1` — Handle существует, `100` — не найден, `2` — ошибка на стороне сервера Handle, пустой ответ — сервер не ответил.

Устойчив ли адрес на CKAN: адрес по UUID должен перенаправлять на адрес по имени.

```
PORTAL=https://data.gov.gr
ID=$(curl -s "$PORTAL/api/3/action/package_search?rows=1" | jq -r '.result.results[0].id')
curl -s -o /dev/null -w '%{http_code} %{redirect_url}\n' "$PORTAL/dataset/$ID"
```

Три проверки на выборке

Проверки DataCite сделаны 12 октября 2026 года; для реестра использован выпуск 1.27.0 от 15 октября.

Ведут ли DOI наборов данных на живые страницы. Случайная выборка DOI ничего не говорит о практике: половина в ней от одного японского института, который регистрирует каждый эксперимент. Поэтому взяты 700 случайных репозиториев DataCite, у 314 из них нашлись наборы данных, и от каждого проверен один случайный DOI набора.

Результат	Репозиториев
страница открылась	233
защита от роботов (403, проверка браузера) — программой не проверить	55
ссылка мертва: «не найдено», несуществующий домен, частный адрес	16
ошибка сервера или нет ответа дважды подряд	10

Среди мёртвых ссылок — DOI, которые с самого начала вели в никуда: на частный адрес `10.4.0.146`, на IP-адрес с портом 8080, на тестовый стенд `nhsdigital-dev.collibra.com` корпоративного каталога. Остальные — следы переездов: «не найдено» у старых записей B2SHARE и у страниц геопортала NOAA 2018 года, несуществующие домены у четырёх университетских репозиториев. Каждый двадцатый репозиторий с наборами данных отдаёт хотя бы часть DOI, которые не открываются, а ещё у каждого шестого программа не может это проверить.

Разрешаются ли Handle в DSpace. Из реестра взяты 120 случайных действующих установок DSpace с OAI-PMH, ответили 86. По идентификатору первой записи проверено, разрешается ли её Handle через hdl.handle.net. У 51 установки — да. У 12 стоит заглушка `123456789` из поставки, у 7 — буквенный локальный префикс вроде `item` или `edata`, который вне сайта ничего не значит. Ещё у 16 префикс числовой, но Handle не найден или сервер Handle не отвечает. Итого у 35 установок из 86 постоянный идентификатор на странице записи — только видимость.

Сколько геокаталогов на деле выдают DOI. Сравнение реестра и DataCite описано выше: 555 установок GeoNetwork с пометкой «DOI» против примерно 45 установок, на которые DOI действительно ведут.

Что из этого следует

- Заложите постоянный адрес до запуска: в адресе набора — неизменный ключ, а для ссылок — адрес, который не зависит от названия.
- Таблицу «старый адрес → новый» и перенаправления готовьте до переезда, а не после. DOI спасает только те наборы, которым его выдали заранее.
- Выдавайте DOI, когда на данные ссылаются как на источник и вы готовы годами держать посадочные страницы. Ежедневно меняющемуся набору DOI не нужен.
- Если выдаёте DOI, выдавайте его неизменным версиям, а при снятии данных оставляйте страницу-надгробие: DOI не удаляется.

- Проверьте адреса в своих DOI и Handle программой: не тестовый стенд, не `localhost`, не заглушка из поставки.
- Расширение для DOI проверяйте по тестам под свою версию ядра. ckanext-doi на CKAN 2.11 и 2.12 не тестировался.
- Помните, что DOI — договор с агентством в чужой юрисдикции. Держите и свой постоянный адрес: он останется, даже если новые DOI получить не удастся.

Источники

- DataCite: состояния DOI, страницы-надгробия, посадочные страницы, версии, взносы, REST API — запросы 12 октября 2026 года: 138,3 миллиона DOI, 76,4 миллиона наборов, 765 DOI на `localhost`, 5145 на общую страницу-заглушку, 7209 на адреса GeoNetwork, выборка по 314 репозиториям
- Crossref: соблюдение санкций, исключение членов
- Handle.Net Registry: регистрация префикса, ARK Alliance, w3id.org, PURL в Интернет-архиве, CSTR
- Dataverse 6.12: поставщики постоянных идентификаторов, версии наборов, цитирование данных
- DOI в InvenioRDM — Crossref с v14, родительский DOI с v12, необязательный DOI с v13, записи с ограниченным доступом
- Версии и DOI в Figshare
- DSpace 10: DOI, префикс Handle по умолчанию в dspace.cfg, поддержка Handle.Net Registry
- GeoNetwork 4.4: настройка DOI, жизненный цикл записи и рабочая копия
- CKAN 2.12: просмотр набора и скачивание ресурса — dataset.py, resource.py; ckanext-doi и его тестовые образы; проверено 12 октября 2026 года
- Huwise: идентификаторы набора в Automation API
- DCAT 3 — свойства версии, серии наборов, контрольная сумма дистрибутива
- Pew Research Center, When Online Content Disappears, май 2024
- T. Berners-Lee, Cool URIs don't change, 1998; политика постоянства адресов W3C
- M. Laakso, L. Matthias, N. Jahn, Open is not forever: a study of vanished open access journals, 2020
- ARDC: национальная стратегия постоянных идентификаторов Австралии, служба DOI DataCite
- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: пометка DOI у 2008 каталогов, выборка 120 установок DSpace с OAI-PMH
- И. Бегтин, Telegram: реестр отечественного ПО и гниение ссылок, 15.04.2021; Crossref, санкции и CSTR, 23.02.2023; австралийская стратегия идентификаторов, 24.03.2023; исчезнувшие журналы открытого доступа, 19.09.2020
- И. Бегтин, Государственные реестры, апрель 2021 — две модели ведения реестров и требования к ним
- Экспорт цитирования, полнота метаданных, жизненный цикл, архивация

- Overture Maps, GERS — реестр идентификаторов, журнал изменений, таблицы соответствия, замена идентификаторов на UUID в выпуске 2025-06-25.0; OpenSanctions, модель утверждений — канонический идентификатор после устранения дублей

Экспорт цитирования

Люди ссылаются на данные, когда это занимает секунду. Если ссылку нужно собирать вручную из названия, года и адреса, большинство процитирует статью или не процитирует ничего.

Главный вывод главы: кнопка «Цитировать» стоит дёшево, дорого то, из чего она собирает ссылку. Генератор за миг соберёт строку из полей описания, но если в них нет автора, года и версии, получится аккуратно оформленная пустышка. Поэтому сначала проверяют поля и то, что набор, на который сослались, через год будет тем же самым. Кнопку делают потом. Если у набора есть DOI, ссылку в десятке форматов уже отдаёт DataCite, и задача каталога — не испортить её своей версией. Проверка в этой главе показывает, что это происходит: для одного и того же набора каталог и DataCite выдают разные ссылки.

Что это

Экспорт цитирования — всё, что помогает сослаться на набор без ручной работы. У него четыре слоя, и платформы закрывают их по-разному.

Строка на странице. Блок «Как цитировать» с готовой ссылкой, которую копируют в текст статьи или отчёта.

Файл для менеджера библиографии. Менеджер библиографии — программа, где исследователь хранит свои источники и откуда вставляет ссылки в текст: Zotero, EndNote, Mendeley, JabRef. Ему отдают ссылку файлом в одном из форматов: BibTeX (для LaTeX), RIS и EndNote XML (для большинства менеджеров), CSL-JSON (для программ, которые оформляют ссылку по стилю CSL). Стиль CSL — описание правил оформления: APA, Chicago, ГОСТ, требования конкретного журнала.

Подсказки для программ на самой странице. Метатеги в заголовке HTML (`citation_title`, `citation_author` и другие, их придумал Google Scholar, или поля Dublin Core). По ним Zotero сохраняет набор одной кнопкой в браузере, а Google Scholar понимает, что перед ним. Сюда же относится заголовок ответа `Link` со ссылкой `rel="cite-as"`: он говорит программе, какой адрес считать адресом для ссылки. Это часть FAIR Signposting, коротко о нём — в глоссарии.

Ссылка по идентификатору, без каталога. Если у набора есть DOI, зарегистрированный в DataCite, его ссылку можно получить у самого DataCite, не заходя на сайт каталога. Этот приём называется согласованием содержимого (content negotiation): программа запрашивает адрес `https://doi.org/…` и в заголовке `Accept` просит не страницу, а, например, BibTeX. Команды — в разделе о проверке.

Две вещи часто путают с экспортом цитирования.

- Свойство `citation` в разметке schema.org — это не ссылка на сам набор, а статьи, которые автор просит процитировать вместе с ним. Google прямо пишет: для ссылки на набор используйте название, идентификатор, автора и издателя, а `citation` — для связанных статей.

- Требование лицензии CC BY указать автора — юридическое условие использования, а не научная ссылка. Принципы цитирования данных FORCE11 различают эти две вещи. Но готовая строка помогает выполнить и то и другое: GBIF, например, по соглашению с пользователем обязывает ссылаться на выгрузку по DOI, а каталог Ifremer пишет в условиях набора, что цитирование обязательно.

Зачем нужна

Чтобы автор данных получил признание. Совместная декларация FORCE11 (2014) сформулировала восемь принципов цитирования данных. Первый — данные считаются полноценным результатом исследования, а ссылка на них весит столько же, сколько ссылка на статью. Для каталога важнее всего три других: уникальный идентификатор, долговечность ссылки и точность — ссылка должна указывать на ту версию или тот срез данных, на которых сделан вывод.

Чтобы ссылку можно было посчитать. Проект Make Data Count вместе с DataCite собирает корпус цитирований данных (Data Citation Corpus). В четвёртом выпуске от 27 июля 2025 года в нём 10,7 млн записей, из них 9,7 млн уникальных пар «набор — публикация». Почти половину записей, 5,2 млн, добавил Europe PMC — собрание статей по наукам о жизни. Его программы находят наборы не в списках литературы, а в тексте статей, где набор упомянут номером записи в базе, например в генетическом банке. Ещё часть корпуса дала такая же программная разметка статей от Chan Zuckerberg Initiative. Значит, на данные ссылаются чаще, чем видно по спискам литературы, но считать такие ссылки трудно и неточно. Аккуратная ссылка с DOI в списке литературы — то, что попадает в учёт без догадок.

Насколько учтённых ссылок мало, видно по самому DataCite. 12 октября 2026 года из его API взяты три тысячи случайных DOI с типом «набор данных». Больше половины из них (1667) выпустил японский Национальный институт термоядерного синтеза (NIFS), поэтому считать лучше без него. Из оставшихся 1333 DOI хотя бы одно цитирование, известное DataCite, есть у 78 — около 6%.

Чтобы владелец каталога видел пользу. Ссылки в статьях и отчётах — самый убедительный довод для руководства и грантодателя, что каталог нужен. Счётчик скачиваний говорит, что файл забрали, ссылка — что из него что-то сделали.

Чтобы пользователь не ошибся. Без готовой строки каждый собирает ссылку сам: кто-то пишет название портала вместо издателя, кто-то — дату скачивания вместо года публикации, почти никто не пишет версию. По такой ссылке набор не найти, а если найти, то не тот.

Из чего состоит ссылка

DataCite рекомендует такой порядок элементов:

```
Автор (Год). Название. Версия. Издатель. Тип ресурса. Идентификатор
```

Минимальная ссылка — без версии и типа: автор, год, название, издатель, идентификатор. Так выглядит ссылка на реальный набор Harvard Dataverse в стиле APA:

```
Durbin, P. (2017). Open Source at Harvard (Version 5.0) [Dataset]. Harvard Dataverse.
https://doi.org/10.7910/DVN/TJCLKP
```

Каждый элемент ломается по-своему.

Автор. Люди или организация. Для государственных данных автор-организация — норма: «Росстат (2026)» лучше, чем фамилия сотрудника, который выгрузил файл. Ошибка — смешать роли: поставить в авторы отдел, который ведёт каталог, или издателя рядом с людьми.

Год. Год публикации набора или его версии, а не год, когда пользователь его скачал, и не год создания карточки. Если программа берёт год из даты заведения записи, после переноса наборов на новый портал у всех них станет год переноса.

Название. Название набора, а не имя файла и не название портала.

Версия. Без неё ссылка ведёт на набор вообще, а не на то, что использовали. В той же выборке DataCite версия заполнена только у 191 DOI из 1333 — у 14%. Но и заполненная версия не всегда доходит до ссылки: BibTeX, который DataCite отдаёт для набора Harvard Dataverse, версии не содержит, хотя у набора пять версий, а ссылка APA от того же DataCite её содержит.

Издатель. Организация, которая опубликовала набор: репозиторий, ведомство, институт.

Тип ресурса. Пометка «[Dataset]» отличает ссылку на данные от ссылки на статью. В классическом BibTeX типа для данных нет, и каждый выкручивается сам: DataCite отдаёт `@misc`, Dataverse — нестандартный `@data`, InvenioRDM — `@dataset`, который понимает biblatex, современная замена BibTeX.

Идентификатор. DOI в виде адреса `https://doi.org/…`. Если DOI нет, ставят постоянный адрес страницы набора и дату обращения. Российский ГОСТ Р 7.0.108-2022 о ссылках на сетевые документы так и требует: адрес и дата обращения обязательны, а DOI в стандарте не упоминается вовсе. Базы данных в нём названы одним из видов электронных документов.

Если поля описания пусты, генератор ссылки не поможет. Какие поля нужны и как проверить, что они заполнены по существу, разобрано в главе о полноте метаданных; как устроены DOI и версии — в главе об идентификаторах.

Ссылка на меняющиеся данные

Научный репозиторий фиксирует версию, и ссылка на неё остаётся точной. Портал открытых данных и каталог показателей чаще ведут один набор, файлы которого заменяют каждую неделю. Ссылка «Минтранс (2026). Расписание пригородных поездов» через месяц указывает уже на другие данные. Способов справиться с этим пять, от простого к сложному.

1. **Дата обращения и архив выгрузок.** Ссылка содержит дату, а портал хранит прошлые файлы, так что состояние на эту дату можно восстановить. Без архива дата обращения только честно признаёт, что проверить ссылку нельзя.
2. **Снимок страницы.** Our World in Data предлагает ссылаться не на живую страницу графика, а на её архивный снимок с датой в адресе (`archive.ourworldindata.org/20260910-…`). Там же он просит сослаться и на

первоисточник, и на себя как на обработчика: «Gapminder (2015); UN IGME (2025) – processed by Our World in Data». Двухуровневая ссылка нужна любому агрегатору и любому порталу, который перепубликует чужие данные. Строку источников программа берёт по соглашению об адресах: к адресу графика дописывают `.metadata.json`, и в ответе есть готовое поле для ссылки, а `.csv` отдаёт сами данные.

3. **Версии с собственным идентификатором.** Каждая выгрузка набора становится версией. Так делают Zenodo и Figshare, так можно делать и на портале, если выпусков немного. CELLxGENE Census добавляет к этому срок. Выпуски с долгой поддержкой обещаны раз в полгода, каждый хранится не меньше пяти лет, а для любого среза есть генератор ссылки. Срок хранения, названный заранее, — то, чего обычно не хватает обещанию «версии сохраняются».
4. **Ссылка на запрос.** GBIF выдаёт свой DOI каждой выгрузке наблюдений: в ней записаны условия отбора, список наборов-источников и дата. 15 октября 2026 года один только набор наблюдений iNaturalist входил в 2 623 241 такую выгрузку. Если пользователь потом сильно отфильтровал выгрузку или собрал данные через API, он регистрирует «производный набор» (derived dataset) и тоже получает DOI. Общая схема описана в рекомендациях Research Data Alliance о ссылках на меняющиеся данные: данные хранят с версиями и метками времени, сохраняют сам запрос с меткой времени и выдают ему идентификатор, по которому запрос можно выполнить снова на том же состоянии данных. Новая редакция рекомендаций одобрена в июле 2025 года: в ней 12 пунктов вместо 14 в редакции 2015 года, суть не изменилась.
5. **Ссылка на часть набора.** Dataverse показывает отдельную ссылку для каждого файла набора, GBIF — для отдельного наблюдения. Это нужно, когда вывод опирается на один файл из ста.

Для портала открытых данных разумный минимум — первый способ: дата обращения в предлагаемой строке и архив прошлых выгрузок. Как хранить прошлые файлы, разобрано в главе о хранилище данных.

Варианты реализации

Шаблон строки на странице. Самый дешёвый способ: в шаблон страницы набора вставляют строку, собранную из полей. Так работает расширение ckanext-doi для CKAN: блок «Cite this as» собирается из автора, года, названия и издателя, к нему добавляется ссылка на DOI. Цена — одна правка темы оформления. Риск — один стиль на всех и ошибки в выборе полей: тот же ckanext-doi берёт год из даты создания записи в CKAN.

Генератор со стилями CSL. Платформа хранит стили CSL и по запросу оформляет ссылку в нужном стиле и формате. Стилей много: в репозитории CSL в октябре 2026 года 2863 самостоятельных стиля и ещё 8002 «зависимых» — для журналов, которые используют чужой стиль под своим именем. Так устроены Dataverse, InvenioRDM и Zenodo, GeoNetwork, расширение ckanext-citeproc для CKAN. Цена — сопровождение набора стилей и соответствия полей переменным CSL.

Ссылка от DataCite. Каталог ничего не генерирует сам, а берёт ссылку у DataCite по DOI. Так делает NADA, а кнопка BibTeX в ckanext-doi ведёт прямо в API DataCite. Плюс — ссылка совпадает с той, что получит любой другой сервис. Минусы — работает только для наборов с

DOI, DataCite ничего не знает о версиях, которые не записаны в его метаданных, и у его генератора есть ловушки:

- неизвестное имя стиля DataCite молча заменяет на APA. Запрос стиля `gost-r-7-0-100-2018` возвращает обычную ссылку APA без всякого предупреждения. InvenioRDM на такой же запрос честно отвечает ошибкой;
- стилей по ГОСТ в репозитории CSL три, все по старому ГОСТ Р 7.0.5-2008, и ни один не выводит для набора данных DOI. Для набора Harvard Dataverse они дают «Durbin P. Open Source at Harvard // 2017.» и «Durbin P. Open Source at Harvard. Harvard Dataverse, 2017.» Если каталогу нужна ссылка по ГОСТ, шаблон для данных лучше написать самому.

Метатеги для менеджеров библиографии. Метатеги `citation_*` и Dublin Core в заголовке страницы стоят почти ничего и дают сохранение одной кнопкой в Zotero и правильную карточку в Google Scholar. В ядре DSpace они есть даже там, где нет никакого блока ссылки. У Zotero есть и отдельные программы-трансляторы для крупных платформ — Dataverse, InvenioRDM, Figshare, Dryad, DSpace. Но транслятор привязан к списку адресов: транслятор InvenioRDM знает 14 сайтов, от zenodo.org до репозитория Тюбингенского университета. Новую установку он не узнает, и Zotero останется только читать метатеги. У метатегов есть предел: в словаре Google Scholar нет ни версии, ни типа «набор данных», поэтому Zotero по ним часто сохраняет набор как веб-страницу или статью. Тип умеет передать старый формат COinS, но его почти никто не ставит. Тип «Dataset» в самом Zotero есть с версии 6.0.24.

Файл в репозитории. Для данных на GitHub роль кнопки играет файл `CITATION.cff` в корне репозитория. GitHub показывает по нему ссылку «Cite this repository» в двух форматах — APA и BibTeX. Чтобы ссылка была на данные, а не на программу, в файле нужна строка `type: dataset`. Без неё GitHub оформит набор как программу: «[Computer software]» в APA и `@software` в BibTeX.

Вариант	Что получает пользователь	Цена	Главный риск
Шаблон строки	одну строку в одном стиле	правка темы	неверно выбранные поля, особенно год
Генератор CSL	строку в любом стиле и файлы для менеджеров	платформа или расширение, набор стилей	плохое соответствие полей переменным CSL
Ссылка от DataCite	строку и файлы, совпадающие с другими сервисами	ничего, если DOI уже есть	нет DOI — нет ссылки; молчаливая замена стиля
Метатеги	сохранение одной кнопкой в Zotero	несколько строк в шаблоне	нет версии и типа «набор данных»
`CITATION.cff`	кнопку на GitHub, метаданные для Zenodo	один файл	без `type: dataset` получится ссылка на программу

Варианты не исключают друг друга. Хороший научный репозиторий делает все сразу: строку, файлы, метатеги и DOI, по которому DataCite отдаёт ту же ссылку.

Когда нужна, а когда нет

Ситуация	Что нужно	Чем обычно закрывают
Научный репозиторий	строка с версией и DOI, BibTeX и RIS, метатеги, ссылка на файл набора	ядро Dataverse, InvenioRDM, Figshare
Портал открытых данных при университете или министерстве, где данными пользуются исследователи	строка с постоянным адресом и датой обращения, архив прошлых выгрузок	правка шаблона; для наборов, на которые ссылаются в статьях, — копия в научном репозитории с DOI
Каталог показателей и статистики	ссылка на ряд или запрос с датой, а не на страницу таблицы	ссылка на сохранённый запрос, как у GBIF; снимок, как у Our World in Data
Геопортал с наборами, на которые ссылаются учёные	строка из полей ISO 19115 и DOI	ядро GeoNetwork 4.4
Хаб для машинного обучения	DOI на ревизию набора и строка в карточке	DOI по запросу на Hugging Face; ссылка от DataCite
Публикация в Git-репозитории	`CITATION.cff` с `type: dataset`, DOI выпуска	GitHub и Zenodo
Служебная витрина, слой внутренней карты	постоянный адрес	генератор ссылок не нужен

Обязательно это научному репозиторию: цитирование — часть его задачи, а журналы и грантодатели требуют ссылку на данные. Желательно везде, где данными пользуются в исследованиях и аналитике. Не нужно, если на данные не ссылаются: оперативная витрина, внутренний слой карты. Даже там постоянный адрес полезен, а генератор BibTeX — нет.

Коротко, как выглядит обоснование по шаблону карточки решения:

- **Контекст.** Портал открытых данных университета на CKAN 2.11. Исследователи просят «кнопку цитирования». DOI у портала нет, договора с DataCite тоже.
- **Варианты.** Строка в шаблоне страницы набора; расширение ckanext-citeproc со стилями CSL; ckanext-doi с договором и префиксом DataCite; отправлять наборы, на которые ссылаются в статьях, в Dataverse или Zenodo.
- **Критерии.** Точность ссылки (версия, постоянный адрес), стоимость обновлений CKAN, нужен ли DOI.
- **Решение.** Строка в шаблоне с постоянным адресом и датой обращения, архив прошлых выгрузок. Наборы к статьям — в научный репозиторий с DOI.
- **Почему.** Десять стилей CSL не делают ссылку без версии и DOI точнее. Строка в шаблоне не зависит от чужого расширения, а ckanext-doi требует договора, ради которого портал не создавался.
- **Когда пересмотреть.** Если университет получит префикс DOI через библиотечный консорциум или если ckanext-citeproc выпустит стабильную версию с проверкой на CKAN 2.12.

Как это сделано в платформах

Пометки — ядро, расширение, нет, не проверено, как в сводной таблице. Сверено с документацией, исходным кодом и живыми установками 12 октября 2026 года.

Платформа	Строка на странице	Файлы и стили	Подсказки для программ	Что важно знать
Dataverse 6.x	ядро: с версией и контрольной суммой содержимого (UNF)	ядро: EndNote XML, RIS, BibTeX, CSL-JSON для набора и для отдельного файла; через API — для любой версии	ядро: метатеги Dublin Core, JSON-LD, заголовок `cite-as`	DOI один на набор, версия — в строке ссылки
InvenioRDM v14, Zenodo	ядро	ядро: шесть стилей в меню (APA, Harvard, MLA, Vancouver, Chicago, IEEE), через API — любой стиль CSL; BibTeX, CSL-JSON	ядро: метатеги `citation_*`, заголовок `cite-as`	у каждой версии свой DOI и общий родительский DOI. RIS и EndNote в API нет
Figshare	ядро: кнопка «Cite»	ядро: RefWorks, BibTeX, EndNote, DataCite, NLM, Dublin Core, RIS	не проверено: страница отвечает роботам проверкой	у каждой версии DOI с суффиксом `.v1`, `.v2`
DSpace 10.1	нет	нет	ядро: метатеги `citation_*` для Google Scholar	кнопку «Cite» переносят из DSpace-CRIS, запросы на перенос в октябре 2026 года открыты
GeoNetwork 4.4	ядро: блок в карточке, включается в настройках интерфейса	ядро: HTML, текст, RIS, BibTeX через API	не проверено	состав строки задан в файле XSL ядра; роли авторов и издателя настраиваются параметрами
NADA	ядро, через DataCite, если у исследования есть DOI	ядро, через DataCite	не проверено	без DOI ссылки нет
CKAN 2.11–2.12	расширение: ckanext-doi (только при DOI), ckanext-citeproc	расширение: BibTeX через DataCite (ckanext-doi), стили CSL (ckanext-citeproc)	нет: метатегов в ядре нет, разметку schema.org даёт ckanext-dcat	расширения молодые или привязаны к DOI, см. ниже
Huwise (OpenDataSoft)	нет	нет	нет	есть текстовые поля для указания источника и поле `bibliographicCitation` в шаблоне DCAT — их заполняют руками
Socrata, ArcGIS Hub	нет	нет	нет	генератора ссылки нет
Hugging Face	строка в карточке набора, если её написал автор	ссылка от DataCite, если выпущен DOI	не проверено	DOI выпускают по запросу на текущую ревизию; при новой ревизии выпускают новый, старый помечается устаревшим
GitHub	кнопка «Cite this repository» по `CITATION.cff`	APA и BibTeX	—	не платформа каталога; нужен `type: dataset`

Dataverse прямо называет форматы цитирования штатной функцией, и его ссылка точнее всех: в ней есть номер версии и контрольная сумма содержимого, по которой можно проверить, что таблица не изменилась. У InvenioRDM ссылка проще, но версию фиксирует сам DOI: у каждой версии он свой.

Расширения и их риски

У CKAN в ядре цитирования нет. Кандидатов три, и у каждого своя оговорка. Методика оценки расширения — в главе об экосистемах расширений.

- **ckanext-doi** Музея естественной истории в Лондоне. Выдаёт DOI через DataCite и показывает блок «Cite this as» с кнопкой BibTeX, которую отдаёт DataCite. Без DOI цитирования нет. Документация называет CKAN 2.9 и 2.10, на 2.11 и 2.12 расширение не тестировалось; подробнее — в главе про CKAN.

- **ckanext-citeproc** выложила команда портала open.canada.ca в марте 2025 года, лицензия MIT. Оформляет ссылку на набор и на ресурс по стилям CSL, которые администратор кладёт в папку на сервере: в комплекте библиотеки только один стиль, Harvard. Совместимость заявлена с CKAN 2.10 и 2.11, автоматические тесты идут на основной ветке CKAN. На самом open.canada.ca (CKAN 2.10.8) в октябре 2026 года расширение не включено.
- **ckanext-citation** группы TIB в Ганновере появилось в июне 2026 года, лицензия AGPL-3.0. Отдаёт BibTeX, RIS, EndNote, APA, формат ACM и CSL-JSON. Заявлены CKAN 2.9 и 2.10. Тесты проверяют только оформление ссылки, на живом CKAN расширение автоматически не проверяется. Под тем же именем есть старое расширение тайваньского репозитория depositar: оно не обновлялось с 2022 года и знает только название, автора и адрес.

Для DSpace 6 существовали сторонние дополнения — модуль CSL компании Atmire и выгрузка в RIS Университета Уаикато. В DSpace 7 и новее они не перенесены. В GeoNetwork состав строки правят в файле XSL внутри схемы ядра: такую правку придётся повторять после каждого обновления.

Как ломается: один набор — три ссылки

Проверка 12 октября 2026 года: для каждого набора сравнивалась ссылка, которую выдаёт сам каталог, и ссылка, которую по тому же DOI отдаёт DataCite.

Harvard Dataverse (Dataverse 6.10.1), набор «Open Source at Harvard». У набора пять версий: первая — июль 2017 года, пятая — июнь 2026 года.

- Строка на странице: `Durbin, Philip, 2017, "Open Source at Harvard", https://doi.org/10.7910/DVN/TJCLKP, Harvard Dataverse, V5, UNF:6:…` — с версией и контрольной суммой.
- BibTeX от самого Dataverse для первой версии: `version = {V1}`, тип `@data`.
- BibTeX от DataCite: версии нет, тип `@misc`.
- APA от DataCite: «(Version 5.0)».

Пользователь, взявший BibTeX у DataCite, сошлётся на набор без версии, то есть фактически на последнюю. Через год это будут уже другие данные.

Каталог Ifremer Sextant (GeoNetwork 4.4.13), климатология температуры и солёности Средиземного моря.

- Строка от каталога: `IFREMER / IDM/SISMER, Simona Simoncelli, Paolo Oliveri . SeaDataCloud Mediterranean Sea - Temperature and Salinity Climatology V1. https://doi.org/10.12770/…` — года нет, организация стоит первым автором.
- BibTeX от каталога: `author = {IFREMER / IDM/SISMER, Simona Simoncelli, Paolo Oliveri}`, `publisher = {}`. BibTeX разделяет авторов словом `and`, а не запятыми, поэтому менеджер библиографии прочитает трёх авторов как одного с длинным именем. Издатель пуст.
- APA от DataCite: `Simoncelli, S., & Oliveri, P. (2019). … (Version 1.1) [Dataset]. IFREMER / IDM/SISMER.` — с годом, версией и издателем.

Поля в описании есть, но генератор каталога берёт их не из тех ролей. Это типичная поломка: функция «в ядре», а результат хуже, чем у внешнего сервиса.

InvenioRDM (демонстрационный стенд) и стиль ГОСТ. BibTeX и APA собраны верно: автор, год, версия, издатель, DOI. Стиль `gost-r-7-0-5-2008` даёт «Gallagher J. и др. Jones and Sons's gallery // 1987–2018.» — без издателя и без DOI. Дело не в платформе, а в стиле CSL: для набора данных он не написан.

Ещё четыре поломки встречаются часто:

- **Нет версии.** Ссылка верна по форме и бесполезна для проверки. В выборке DataCite версия заполнена у 14% наборов.
- **Год из даты карточки.** Программа подставляет год создания записи, а не публикации данных. После переноса на новую платформу все наборы «публикуются» в год переноса.
- **Молчаливая подмена стиля.** DataCite на неизвестный стиль отдаёт APA, и пользователь уверен, что получил то, что просил.
- **Ссылка на программу вместо данных.** `CITATION.cff` без `type: dataset` на GitHub даёт «[Computer software]».

Публикация без портала

Без портала ссылку целиком собирает автор, зато и ошибиться в ролях некому, кроме него. Хватает трёх вещей.

`CITATION.cff` в корне репозитория. Версия формата — 1.2.0, она не менялась с 2021 года. Обязательны `cff-version`, `message`, `title` и `authors`; для данных добавляют `type: dataset`, версию, дату выпуска и DOI, когда он появится. GitHub покажет по файлу кнопку «Cite this repository». Zenodo при архивации выпуска возьмёт из него авторов и название, но только если в репозитории нет файла `.zenodo.json`: если он есть, `CITATION.cff` Zenodo игнорирует. Тип записи Zenodo — набор данных, а не программа — задают полем `upload_type` в `.zenodo.json`. Как совместить оба файла, разобрано в главе о Git-репозиториях. Проверить файл можно программой `cffconvert --validate`. Готовый шаблон CITATION.cff лежит в шаблонах книги.

Строка «Как цитировать» в README и на странице набора. Одна строка в стиле, принятом в вашей области, с версией и постоянным адресом, а после архивации в Zenodo — с DOI. У Zenodo две ссылки на выбор: DOI конкретного выпуска и общий DOI всех версий, который всегда ведёт на последнюю. Для статьи нужен первый.

Разметка на странице набора. Разметка schema.org с автором (`creator`), датой публикации (`datePublished`), версией (`version`) и издателем (`publisher`) — её читают Google и агрегаторы. Поле `identifier` добавляют, когда DOI уже выдан: в шаблоне его нет, чтобы не опубликовать пустой номер. Для Zotero полезны ещё несколько метатегов:

```
<meta name="citation_title" content="Название набора данных">
<meta name="citation_author" content="Фамилия, Имя">
<meta name="citation_date" content="2026/01/15">
<meta name="citation_publisher" content="Организация-издатель">
```

Когда DOI выдан, к этим полям добавляют `citation_doi` с настоящим номером. Версию в метатеги не передать, поэтому она должна быть в названии или в строке «Как цитировать». Как устроить сам репозиторий и выпуски, разобрано в главе о Git-репозиториях.

Как проверить живой портал

Ручная проверка — пять вопросов к пяти случайным наборам:

1. На странице набора есть блок «Как цитировать» с автором, годом, названием, издателем и идентификатором.
2. Год в ссылке — год публикации набора или версии, а не год создания карточки и не сегодняшний.
3. Для двух версий одного набора ссылки различаются номером версии или идентификатором. Если версий нет, в ссылке есть дата обращения.
4. Скачивается BibTeX или RIS, в нём тот же идентификатор и та же версия, что на странице, а авторы разделены как положено (в BibTeX — словом `and`).
5. Если у набора есть DOI, ссылка от DataCite совпадает со ссылкой на странице по автору, году и версии.

Команды

Ссылка по DOI от DataCite — в стиле, в BibTeX и в RIS. Работает для любого DOI, выпущенного через DataCite, даже если у каталога нет никакой кнопки:

```
DOI=10.7910/DVN/TJCLKP
curl -sLH "Accept: text/x-bibliography; style=apa" "https://doi.org/$DOI"; echo
curl -sLH "Accept: application/x-bibtex" "https://doi.org/$DOI"; echo
curl -sLH "Accept: application/x-research-info-systems" "https://doi.org/$DOI"
```

Если стиль написан с ошибкой, ответ всё равно придёт — в стиле APA. Проверяйте, что получили то, что просили. Список имён стилей — в репозитории CSL.

Dataverse: ссылка для конкретной версии в BibTeX, RIS, EndNote или CSL-JSON:

```
SERVER=https://dataverse.harvard.edu
PID=doi:10.7910/DVN/TJCLKP
curl -s "$SERVER/api/datasets/:persistentId/versions/1.0/citation/BibTeX?persistentId=$PID"
curl -s "$SERVER/api/datasets/:persistentId/versions/:latest-published/citation?
persistentId=$PID"
```

InvenioRDM: строка в выбранном стиле и BibTeX. Пример — CaltechDATA; сам Zenodo программам без браузера в октябре 2026 года не отвечал:

```
SERVER=https://data.caltech.edu
ID=$(curl -s "$SERVER/api/records?size=1&q=metadata.resource_type.id:dataset" | jq -r
'.hits.hits[0].id')
curl -s -H "Accept: text/x-bibliography" "$SERVER/api/records/$ID?style=apa&locale=en-US";
echo
curl -s -H "Accept: application/x-bibtex" "$SERVER/api/records/$ID"
```

GeoNetwork 4.4: список форматов и ссылка в RIS для записи:

```
API=https://sextant.ifremer.fr/geonetwork/srv/api
UUID=ad07a55f-5de7-4abc-ba89-8899b16c4b59
curl -s "$API/records/$UUID/formatters/citation?format=?"; echo
curl -s "$API/records/$UUID/formatters/citation?format=ris"
```

Подсказки для программ на странице — метатеги и заголовок `cite-as`:

```
PAGE=https://dspace.mit.edu/handle/1721.1/93094
curl -sL "$PAGE" | grep -o -E '<meta name="(citation_|DC\.)[a-zA-Z_.]*"' | sort | uniq -c
curl -sI https://data.caltech.edu/records/$ID | grep -i '^link' | tr ',' '\n' | grep cite-as
```

Пустой вывод первой команды значит, что Zotero сохранит набор как обычную веб-страницу. Пустой вывод второй — что программа не узнает, какой адрес считать адресом для ссылки.

Что из этого следует

- Сначала поля, потом кнопка: автор, год публикации, название, версия, издатель, идентификатор. Без них любой генератор выдаёт пустышку.
- Если у наборов есть DOI, сверьте свою ссылку со ссылкой DataCite. Расхождение по автору, году или версии — ошибка, которую исправляют в описании, а не в шаблоне.
- Для меняющихся данных добавьте в ссылку дату обращения и храните прошлые выгрузки. Для научных ссылок — фиксируйте версии или выдавайте идентификатор выгрузке, как GBIF.
- Метатеги `citation_*` и Dublin Core стоят несколько строк шаблона и экономят пользователю ручной ввод.
- Ссылку по ГОСТ для набора данных стили CSL не соберут правильно: напишите шаблон сами.
- В Git-репозитории с данными нужен `CITATION.cff` со строкой `type: dataset`.

Источники

- Joint Declaration of Data Citation Principles, FORCE11, 2014 — восемь принципов, различие юридической и научной атрибуции
- DataCite: согласование содержимого и форматы, рекомендуемый порядок элементов ссылки — по руководству CESSDA; выборка 3000 случайных DOI с типом «набор данных» через API DataCite, 12 октября 2026 года
- Make Data Count: документация Data Citation Corpus, данные выпуска 4.0, 27 июля 2025 года
- Research Data Alliance: Precise, Actionable Reference to Dynamic Data, 2025; редакция 2015 года
- Правила цитирования GBIF — DOI выгрузок и производные наборы; Our World in Data: как цитировать
- Разметка Dataset в документации Google — свойство `citation` не для ссылки на сам набор
- Dataverse: возможности, API цитирования, цитирование данных
- InvenioRDM: сериализаторы записей, стили по умолчанию

- Figshare: цитирование и выгрузка ссылок
- DSpace: метатеги для Google Scholar; перенос выгрузки ссылок из DSpace-CRIS — задача 11783, запросы 11451 и dspace-angular 4779, открыты на 12 октября 2026 года
- GeoNetwork 4.4: цитирование записи
- CKAN: ckanext-doi, ckanext-citeproc, ckanext-citation (TIB), ckanext-citation (depositar)
- Huwise: поля для совместимости, Hugging Face: DOI
- GitHub: файлы CITATION; Citation File Format; Zenodo: CITATION.cff и .zenodo.json
- Zotero: как отдать метаданные, тип Dataset, трансляторы
- Репозиторий стилей CSL, подсчёт на 12 октября 2026 года; biblatex: тип dataset
- ГОСТ Р 7.0.108-2022 — библиографические ссылки на электронные документы в сетях
- Живые проверки 12 октября 2026 года: Harvard Dataverse, набор Open Source at Harvard, Ifremer Sextant, климатология Средиземного моря, демонстрационный стенд InvenioRDM, CaltechDATA, DSpace@MIT
- Идентификаторы и DOI, полнота метаданных, Git-репозитории
- Our World in Data, Chart API — суффиксы `.csv`, `.metadata.json`, `.zip`; CZ CELLxGENE, Census — выпуски с долгой поддержкой раз в полгода, хранение не меньше пяти лет; GBIF, выгрузки с набором iNaturalist — 2 623 241 на 15 октября 2026 года

Доступ, роли и процессы

Каталог на одного публикатора живёт без ролей: есть человек, который выкладывает файлы. Каталог организации почти сразу становится общим. Тогда нужно договориться, кто создаёт набор, кто его проверяет, кто видит файл и кто вообще может завести учётную запись.

Практический вывод главы такой. Большинству государственных порталов хватает двух ролей — администратор и редактор организации — и черновика, который не виден в поиске. Проверка перед публикацией нужна, когда публикуют люди, которым вы не доверяете выложить данные сразу: внешние авторы научного репозитория, сотни муниципалитетов, пользователи хаба. Ограниченный доступ нужен персональным, лицензируемым и чувствительным данным, и при этом описание набора остаётся открытым. Открытая регистрация нужна только там, где публикуют внешние люди, и тогда — с приглашением или проверкой, а не для всех подряд. Читать описания и скачивать открытые файлы учётная запись не должна требовать вовсе.

В конце главы — замер на 1125 порталах CKAN 12 октября 2026 года: у скольких открыта регистрация, сколько из них уже заполнено спамными учётными записями и какие расширения входа и доступа стоят на самом деле.

Что это

Под словами «доступ и роли» прячутся пять разных вопросов. Справочник научных репозиториев re3data разделяет по крайней мере три из них: кто может зайти в репозиторий, кто может скачать данные и кто может их загрузить. Ответы на них не обязаны совпадать: в Zenodo загружает любой зарегистрированный, а скачать закрытый файл можно только с разрешения автора.

Кто публикует. Организация-издатель, редактор, куратор, администратор. **Роль** — набор прав, который выдают человеку в пределах организации или коллекции: «редактор организации А» может править наборы А и не может трогать наборы Б. Права дают роли, а не отдельному человеку, чтобы при смене сотрудника не перебирать сотню галочек.

Что происходит до публикации. Черновик — набор, который уже заведён, но не виден в поиске и в выгрузке. **Модерация**, или проверка перед публикацией, — шаг, на котором другой человек (куратор, модератор, администратор организации) смотрит набор и выпускает его или возвращает автору. **Журнал изменений** записывает, кто, когда и что поменял в описании.

Кто видит данные. Открытый файл, файл после принятия условий, после регистрации, по одобренной заявке, после даты эмбарго, только в защищённой среде или вовсе никому — только описание. Ограничить можно всю запись, только файлы или отдельный файл, а у облачных сервисов — даже отдельные колонки и строки таблицы.

Как входят. Здесь два термина, которые часто путают. **Аутентификация** — проверка, кто вы: пароль, второй фактор, государственная учётная запись. **Авторизация** — проверка, что вам можно: роль, права на набор, одобренная заявка. **Единый вход** (SSO, single sign-on)

переносит аутентификацию из каталога во внешнюю систему организации: сотрудник входит той же учётной записью, что и в почту. Для этого используют два стандарта. **SAML 2.0** — стандарт OASIS 2005 года, на нём работают университетские федерации и корпоративные каталоги пользователей. **OpenID Connect** — слой проверки личности поверх протокола OAuth 2.0; его поддерживают Keycloak, облачные поставщики учётных записей и многие государственные системы входа.

Кто может завести учётную запись. Регистрация бывает закрытой (учётные записи заводит администратор или приглашают по почте), открытой с проверкой (заявку одобряет администратор, разрешены только адреса определённых доменов), открытой с капчей и открытой для всех.

Зачем нужна

Чтобы ведомства не правили чужое и было видно, кто отвечает. Когда на портал публикуют двадцать организаций, каждая должна видеть и менять только свои наборы. Роль заодно отвечает на вопрос пользователя «кто за это отвечает»: у набора видна организация и контакт, а не только имя сайта. Об издателе и организации в описании — в главе «Анатомия каталога».

Чтобы ошибку поймали до публикации. Опубликованный набор за часы забирают сборщики, поисковики и веб-архивы. Файл с необезличенными персональными данными, опубликованный по ошибке, потом почти невозможно убрать из всех копий. Черновик и второй взгляд перед публикацией дешевле, чем разбор утечки.

Чтобы показать данные, которые нельзя открыть. Микроданные обследований, медицинские записи, лицензируемые базы нельзя выложить всем, но о них нужно знать. В ЕС это уже обязанность. По Акту об управлении данными (регламент 2022/868, применяется с 24 сентября 2023 года) каждая страна держит единую информационную точку с перечнем охраняемых данных государственного сектора: формат, размер и условия повторного использования. Сами данные выдают по запросу, в том числе только в защищённой среде обработки. Перечни стран собирает европейский реестр охраняемых данных (European Register for Protected Data, ERPD) на data.europa.eu. Регламент о Европейском пространстве медицинских данных (2025/327) требует от органов доступа к медицинским данным открытого машиночитаемого каталога наборов с условиями доступа; эта часть применяется с 26 марта 2029 года.

Чтобы портал не заняли роботы. Открытая регистрация на публичном сайте — приглашение для спама. Роботы заводят учётные записи и от их имени создают организации, группы, наборы и запросы данных с рекламой казино и мошенническими ссылками. Так в 2013 году пришлось закрыть регистрацию на DataHub.io, в 2021 году спамом было забито демонстрационное demo.ckan.org, а у портала метеослужбы Туниса в 2025 году было больше 1,3 миллиона учётных записей и больше 45 тысяч спам-текстов в разделе запросов данных. Разработчики CKAN поэтому выключили открытую регистрацию по умолчанию: в 2.9.8, а в линейке 2.10 — только с 2.10.4. В 2.10.0–2.10.3 она по умолчанию снова открыта. Замер в конце главы показывает, что проблема не историческая.

Чтобы знать публикатора в лицо. На государственных порталах публикатор часто входит через государственную систему идентификации, и это не формальность: так портал знает, от имени какого ведомства опубликован набор. На data.gov.ru по методическим рекомендациям 2024 года поставщик — орган власти или местного самоуправления — входит через ЕСИА, его карточку проверяет администратор, публиковать может только проверенный поставщик, и наборы проходят модерацию. На норвежском data.norge.no описания регистрируют после входа через национальную систему ID-porten и только с ролью в реестре полномочий Altinn. На data.gouv.fr сотрудники организаций входят через ProConnect — государственный вход для специалистов на OpenID Connect, рядом с обычной регистрацией по почте.

Обратная сторона — цена. Каждая роль, шаг проверки и вид доступа — это настройка, которую нужно сопровождать, объяснять новым сотрудникам и проверять при обновлении платформы. Модерация, которую некому вести, превращается в очередь из сотни наборов, которые месяцами не выходят.

Варианты реализации

Кто публикует: пять моделей прав

Без ролей. Один человек или одна команда с общими правами. Так живут статические сайты, Git-репозитории и небольшие порталы. Этого хватает, пока публикует одна команда.

Организации с ролями. Набор принадлежит организации, у людей в организации роли. В CKAN их три: участник видит закрытые наборы организации, редактор создаёт, правит, удаляет наборы и переключает их между открытыми и закрытыми, администратор ещё и управляет участниками. Над всеми — системный администратор. В udata, на котором работает data.gouv.fr, роли администратора и редактора, а с версии 15.4.0 (март 2026 года) ещё и частичный редактор. В Socrata роли выдают на весь портал: администратор, публикатор, редактор, дизайнер, наблюдатель. Это самая распространённая модель на порталах открытых данных.

Коллекции с наследованием прав. Права выдают на коллекцию, и они переходят на всё, что внутри. Так устроены Dataverse (коллекции и наборы в них), DSpace (сообщества, коллекции, записи и файлы) и GeoNetwork, где у каждой группы свой набор прав на запись: просмотр, скачивание, показ на интерактивной карте, правка, уведомления. Модель удобна университету: факультет — коллекция, кафедра — коллекция внутри.

Сообщества с кураторами. Автор публикует сам, а сообщество решает, включить ли запись в свою коллекцию. Так работают InvenioRDM и Zenodo: у сообщества есть владелец, менеджеры, кураторы и читатели, а запись попадает в сообщество после проверки куратором.

Права на отдельный набор. Права выдают не по организации, а на каждый набор: в Huwise (бывший OpenDataSoft) — «просмотр», «просмотр и правка», «просмотр, правка и публикация» для пользователя или группы; в ArcGIS Online — уровни доступа «только владелец», группы, организация и все. В CKAN похожую роль играют соавторы набора (collaborators): с версии 2.9 администратор организации может дать отдельному человеку права участника или редактора на один набор. По умолчанию эта возможность выключена.

Чем тоньше права, тем дороже сопровождение. Матрица из двадцати ролей почти никогда не окупается: через год никто не помнит, зачем она так устроена, а ошибки в правах находят по жалобам.

Что происходит до публикации

Публикация сразу. Сохранил — опубликовал. Так по умолчанию работают CKAN и DKAN 2.x: новый набор в DKAN сразу получает состояние «опубликован», если администратор не поменял настройку. Годится, если публикуют свои сотрудники и ошибку можно быстро исправить.

Черновик. Набор сохранён, но не виден в поиске. В CKAN многошаговая форма создания сохраняет набор черновиком, а в 2.12 появились кнопки «Опубликовать» и «Снять с публикации». В udata черновик — это закрытый набор; при работе через API флаг закрытости нужно передавать при каждом запросе, иначе набор станет публичным. В Huwise правки сохраняют отдельно и выпускают повторной публикацией. Черновик ничего не стоит и нужен всем, кроме совсем простых каталогов.

Раздельное право публиковать. Готовить набор может один, выпускать — другой. В Socrata редактор загружает данные, но публикует только администратор, публикатор или владелец набора. В Huwise права «создавать наборы» и «править все наборы» не включают право публиковать. Это ещё не проверка, но уже второй человек перед выходом набора.

Проверка куратором. Автор отправляет набор на проверку, куратор выпускает его или возвращает с замечанием. В Dataverse участник отправляет набор на проверку, куратор коллекции публикует или возвращает автору. В DSpace у коллекции настраивается процесс из нескольких шагов: проверить, отредактировать, окончательно одобрить. В GeoNetwork 4.4 процесс «черновик → на проверке → одобрено → снято» включают для всего каталога или отдельных групп, а правка одобренной записи создаёт рабочую копию, которую не видят посторонние до повторного одобрения. Тонкость GeoNetwork: по умолчанию даже при включённом процессе запись можно опубликовать без одобрения (параметр `allowPublishNonApprovedMd`), и это надо выключить отдельно. В Socrata одобрение перед публикацией включается в настройках, автоматическое или ручное. В CKAN такого шага в ядре нет, только расширения — о них ниже.

Журнал изменений. Кто, когда и что поменял. В CKAN с версии 2.10 журнал вынесен в отдельный модуль `activity`, который входит в поставку, но должен быть включён. Он хранит снимок набора при каждой правке и показывает разницу между версиями; посторонним подробности по умолчанию не видны. В научных репозиториях роль журнала играют версии: правка опубликованного набора в Dataverse создаёт новую версию, прежняя остаётся. О версиях — в главе «Постоянные идентификаторы, DOI и версии».

Часть проверки лучше отдать программе, а не человеку: обязательные поля, лицензия из списка, файл открывается, кодировка верна. Тогда куратор смотрит на содержание, а не на опечатки. Об этом — в главе «Полнота и детальность метаданных».

Кто видит данные: уровни доступа

Уровень	Что делает пользователь	Где встречается
Открыто	скачивает без условий	любой каталог
Открыто после принятия условий	ставит отметку «согласен» или заполняет короткую анкету	«прямой доступ» в NADA, гостевая книга в Dataverse, закрытые наборы с автоматическим одобрением на Hugging Face
После регистрации	входит в систему	файлы общего пользования в NADA
По заявке	подаёт запрос, ждёт решения владельца	Dataverse, InvenioRDM и Zenodo, DSpace (запрос копии), Huwise, Hugging Face с ручным одобрением, «лицензированные данные» в NADA
Эмбарго	ждёт даты, после которой файл откроется	Dataverse, InvenioRDM и Zenodo, DSpace
Защищённая среда	работает с данными на чужом сервере или в специальном помещении, забирает только проверенные результаты	«защищённая среда» в NADA, доверенные исследовательские среды, условие Акта об управлении данными
Только описание	узнаёт, что данные есть и у кого их просить	«данные недоступны» в NADA, реестр охраняемых данных ЕС

Как устроены эти уровни в каталогах микроданных, подробно разобрано в главе о каталогах микроданных.

Два уточнения. Регистрация — ещё не закрытый доступ. Словарь видов доступа Бюро публикаций ЕС считает публичным и набор за регистрацией или ключом API, если их может получить любой. Для открытых данных в смысле этой книги регистрация — уже барьер, но это другой разговор: см. глоссарий, статью «Общедоступные и открытые данные». И «по запросу» без срока ответа и без условий, записанных заранее, — не уровень доступа, а отказ, растянутый во времени.

Описание открыто, файл закрыт

Закрывать описание вместе с файлом — плохой обмен. Скрытый набор нельзя найти, на него нельзя сослаться, и запрос на доступ некому отправить. Принципы FAIR прямо разделяют эти вещи: набор с закрытыми файлами может быть находимым и доступным по понятной процедуре, если его описание открыто и у него есть постоянный идентификатор.

Платформы ведут себя здесь очень по-разному, и это стоит проверить до выбора.

- **Умеют в ядре:** Dataverse, InvenioRDM и Zenodo, DSpace, NADA — описание открыто, закрыты файлы, а пользователь видит, как их получить. В GeoNetwork скачивание — отдельное право, которое можно не давать гостям; оно защищает файлы, загруженные в сам GeoNetwork, но не внешние ссылки. Huwise показывает набор в каталоге всем, а содержимое — только выбранным пользователям и группам, с кнопкой «Запросить доступ»; кнопка работает, только если на портале включён режим маркетплейса.
- **Не умеют:** в CKAN закрытый набор скрыт целиком, вместе с описанием. Открыть описание и закрыть отдельный файл позволяет только расширение ckanext-restricted. В Socrata закрытый набор не появляется в публичном каталоге вообще; можно закрыть лишь отдельные поля описания у открытого набора.

Чтобы закрытый набор нашли не только на сайте, условия доступа должны попасть в выгрузку описаний.

- **DCAT.** Свойство `dct:accessRights` со значением из словаря Бюро публикаций ЕС: PUBLIC, RESTRICTED, NON_PUBLIC (в словаре есть и уточнения SENSITIVE и CONFIDENTIAL внутри NON_PUBLIC). В DCAT-AP 3.0.1 свойство необязательное, но европейский реестр охраняемых данных требует его у каждого набора и запрещает значение «открыто», а в HealthDCAT-AP оно обязательно вместе с указанием органа доступа к медицинским данным.
- **DataCite и OpenAIRE.** Поле прав со значением из словаря COAR: открытый доступ, эмбарго, ограниченный доступ, только описание. Так режим доступа видят OpenAIRE и сборщики научных данных.
- **schema.org.** Свойство `isAccessibleForFree` и текстовое `conditionsOfAccess`. Google Dataset Search читает первое и не требует, чтобы данные были открыты: ему нужна открытая страница с описанием, без входа и без запрета для робота.

О форматах выгрузки — в главе «Экспорт метаданных», о лицензии рядом с условиями доступа — в главе «Лицензии и условия доступа».

Насколько закрытый доступ вообще распространён, показывают два крупных репозитория. 12 октября 2026 года в Zenodo из 726 858 наборов данных были открыты 680 981 (94%), с закрытыми файлами — 41 715 (5,7%), под эмбарго — 4162 (0,6%). В Harvard Dataverse из 5,55 миллиона файлов закрыты 76 579 (1,4%) и ещё 1513 — под эмбарго. Закрытые файлы — исключение даже там, где их легко сделать.

Как входят

Пароль на портале. Работает везде и годится для небольшой редакции. Слабое место — пароли, которые подбирают перебором. Защищает второй фактор: одноразовый код из приложения. В CKAN его даёт расширение ckanext-security, в Huwise и ArcGIS Online его можно сделать обязательным в настройках. Таиланд поставил двухфакторный вход в типовую сборку порталов: в замере расширение `thaigdc2fa` стоит на 160 из 250 тайских серверов CKAN.

Единый вход через систему организации. Публикатор входит учётной записью ведомства или университета по SAML или OpenID Connect. Нужен, когда публикаторов много, они приходят и уходят, а заводить и отключать им отдельные пароли некому. Уволенный сотрудник теряет доступ к порталу вместе с почтой. В Dataverse, InvenioRDM, DSpace, GeoNetwork, Huwise, Socrata и ArcGIS Online единый вход есть в ядре или в настройках сервиса, в CKAN — только расширениями.

Государственная система входа. ЕСИА в России, ID-porten в Норвегии, ProConnect во Франции. Портал получает подтверждённую личность и организацию публикатора, но привязывается к чужой системе: при её сбое не может войти ни один публикатор.

Научная учётная запись. ORCID — идентификатор исследователя, который научные репозитории принимают и как способ входа. Удобен репозиторию, куда приходят авторы из разных организаций.

Ключ для программ. Публикация через API идёт не по паролю, а по ключу (токену), который выдают на программу и отзывают отдельно. В CKAN токены есть с версии 2.9, а входящий в поставку модуль `expire_api_token` ограничивает срок их жизни.

Вход нужен публикатору, а не читателю. Если для просмотра описаний или скачивания открытых файлов портал требует учётную запись, это уже не вопрос безопасности, а дверь, закрытая для программ.

Кто может завести учётную запись

Вариант	Как устроено	Когда годится
Закрытая регистрация	учётные записи заводит администратор или приглашает администратор организации по почте	публикуют только сотрудники известных организаций; обычный случай государственного портала
Открытая с проверкой	заявку одобряет администратор, или регистрация разрешена только с адресов нужных доменов	внешние публикаторы есть, но их немного
Открытая с капчей	любой может зарегистрироваться, капча отсекает простых роботов	научный репозиторий, хаб, площадка с обсуждениями — вместе с проверкой того, что публикуют
Открытая для всех	любой регистрируется и сразу может публиковать	почти никогда для публичного каталога

Спамеру не нужна возможность публиковать наборы. Ему хватает любого места, где текст и ссылка появятся на открытой странице: профиль пользователя, новая организация или группа, запрос данных, комментарий. Поэтому важна не только регистрация, но и то, что может сделать новая учётная запись. В CKAN любой вошедший пользователь по умолчанию может создавать группы и организации (параметры `ckan.auth.user_create_groups` и `ckan.auth.user_create_organizations`), а в расширении запросов данных — создавать запросы и комментарии без проверки. В ArcGIS Hub Premium учётные записи сообщества, которые заводятся открытой регистрацией, по умолчанию получают роль публикатора. В Socrata самостоятельно зарегистрированная учётная запись, наоборот, не получает никакой роли на портале.

Капча помогает, но не решает. CKAN умеет Google reCAPTCHA второй версии: на регистрации, а с версии 2.11 ещё и на входе и сбросе пароля. Лучше работает автоматическая проверка содержимого: udata помечает подозрительные наборы, организации, профили, примеры использования и обсуждения по списку слов и отправляет их модератору.

Насколько это распространено

По выгрузке реестра 1.27.0 ограниченный доступ хотя бы к части данных отмечен у 1131 каталога из 43 614, то есть у 2,6%. У 537 из них открытых данных нет вовсе, у 594 открытое и закрытое живёт рядом. Запросы — в приложении.

Интереснее, на каких платформах эти каталоги работают.

Платформа	Каталогов с ограниченным доступом	Всего в реестре, все типы	Доля
DHIS2 (информационная система здравоохранения)	72	74	97%

Платформа	Каталогов с ограниченным доступом	Всего в реестре, все типы	Доля
Gen3 (платформа биомедицинских данных)	29	30	97%
Redivis (платформа данных для исследований)	12	13	92%
NADA	114	159	72%
Pure (система учёта научной работы)	133	369	36%
InvenioRDM	4	76	5%
Dataverse	5	257	1,9%
DSpace	10	830	1,2%
CKAN	12	1951	0,6%
GeoNetwork	1	764	0,1%

Здесь считаются все каталоги на платформе, какого бы типа они ни были. В профилях платформ числа меньше, потому что там учтён только основной тип: например, у CKAN 1951 каталог всего и 1906 порталов открытых данных, у DSpace 830 и 829 научных репозиториев, у NADA 159 и 158 каталогов микроданных.

Закрыть файл при открытом описании умеют в ядре и Dataverse, и DSpace, и InvenioRDM, но каталогов, где ограниченный доступ — заметная часть содержимого, на них единицы. Каталоги с закрытыми данными работают на платформах, где закрытость — норма: системы здравоохранения, биомедицинские платформы, каталоги микроданных. Вывод для выбора: если ограниченный доступ — главная задача, а не редкое исключение, смотрите на платформы, сделанные под неё, а не на портал открытых данных с расширением. Пометка в реестре относится к каталогу в целом: Harvard Dataverse в реестре открыт, хотя полтора процента его файлов закрыты.

Когда нужна, а когда нет

Ситуация	Что нужно	Чего не нужно
Одна команда публикует свои данные	черновик, закрытая регистрация	ролей и модерации
Ведомство и подведомственные организации, региональный портал с муниципалитетами	организации с ролями администратора и редактора, черновик, журнал изменений, закрытая регистрация с приглашениями	многошаговой модерации: каждая организация отвечает за своё сама
Национальный портал, который собирает описания из других каталогов	права только для своей редакции; публикаторы работают в своих каталогах	учётных записей для сотен публикаторов на национальном портале
Университетский репозиторий, куда загружают сотрудники	коллекции факультетов, проверка куратором библиотеки, единый вход через университет, эмбарго	открытой регистрации для всех
Общий научный репозиторий или хаб для внешних авторов	открытая регистрация с капчей, кураторы сообществ, проверка содержимого на спам, закрытые файлы по запросу	ручного одобрения каждой учётной записи
Каталог микроданных	несколько уровней доступа к файлам, заявки с журналом решений, открытое описание всегда	закрытия описания вместе с файлом

Ситуация	Что нужно	Чего не нужно
Охраняемые данные государственного сектора по Акту об управлении данными	открытое описание с `dct:accessRights`, условия и порядок запроса, срок ответа	хранения самих данных в каталоге
Портал с обсуждениями, запросами данных, примерами использования	модерация пользовательских текстов, ограничение прав новых учётных записей	открытой регистрации без проверки содержимого

Роли нужны, как только публикуют больше одной команды. Модерация нужна, если круг публикаторов шире круга тех, кому вы доверяете выложить данные сразу. Ограниченный доступ нужен персональным, лицензируемым и чувствительным данным. Он не нужен, чтобы «посмотреть, кто скачивает»: для этого есть статистика, а не закрытие файла.

Коротко, как выглядит обоснование по шаблону карточки решения:

- **Контекст.** Портал министерства на CKAN обновили с 2.8 до 2.11. Публикуют 14 подведомственных организаций, около 60 сотрудников. Регистрация открыта с первых лет, учётных записей десятки тысяч, в списке пользователей и разделе запросов данных — реклама. Министерство просит «включить модерацию всех наборов».
- **Варианты.** Оставить как есть; закрыть регистрацию и приглашать публикаторов; поставить расширение проверки перед публикацией; перевести вход на ведомственный единый вход.
- **Критерии.** Сколько людей публикуют и кто за них отвечает; кто будет вести модерацию и сколько времени у него на это есть; совместимость расширений с версией ядра; что увидит пользователь портала.
- **Решение.** Закрыть регистрацию, закрыть список пользователей (`ckan.auth.public_user_details = false`), удалить спамные учётные записи, запретить обычным пользователям создавать организации и группы. Публикаторов приглашают администраторы организаций. Роли — администратор и редактор организации, черновик и журнал изменений включены. Модерацию всех наборов не вводить; модерировать только раздел запросов данных.
- **Почему.** Публикуют сотрудники известных организаций, и за наборы отвечает сама организация. Модерация 14 организаций ложится на одного человека в министерстве и станет очередью. Расширение проверки перед публикацией тестируется только на 2.10 и 2.11, и при переходе на 2.12 портал получит лишнее препятствие.
- **Когда пересмотреть.** Если наборы начнут публиковать внешние организации или граждане, если ошибки в опубликованных наборах будут повторяться, если число публикаторов вырастет настолько, что заводить и отключать их вручную станет некому, — тогда переходить на единый вход.

Как это сделано в платформах

Пометки — ядро, расширение, нет, не проверено, как в сводной таблице. Сверено с документацией и репозиториями 12 октября 2026 года.

Платформа	Роли	Проверка перед публикацией	Описание открыто, файл закрыт	Запрос доступа, эмбарго	Единый вход	Регистрация
CKAN 2.12	ядро: организации, три роли, соавторы набора	черновик — ядро; проверка куратором — расширение	нет: закрытый набор скрыт целиком; расширение ckanext-restricted	расширение (ckanext-restricted); эмбарго нет	расширение: SAML, OpenID Connect, LDAP	закрыта по умолчанию с 2.9.8 и с 2.10.4; приглашение по почте — ядро
Dataverse	ядро: роли на коллекцию и набор, группы	ядро: отправка на проверку куратору	ядро	ядро: запрос доступа, условия, гостевая книга, эмбарго	ядро: Shibboleth (SAML), OpenID Connect, ORCID	настраивается
InvenioRDM, Zenodo	ядро: сообщества с владельцем, менеджерами, кураторами	ядро: проверка при включении в сообщество	ядро: закрыта запись целиком или только файлы	ядро: запросы на доступ, эмбарго	ядро: ORCID, OpenID Connect, SAML	у Zenodo открыта
DSpace 10.1	ядро: группы, права на сообщество, коллекцию, запись и файл	ядро: настраиваемый процесс из нескольких шагов	ядро	ядро: эмбарго, запрос копии (Request a Copy)	ядро: Shibboleth, LDAP, ORCID, OpenID Connect	настраивается
NADA	ядро	не проверено	ядро: семь типов доступа к файлам	ядро: заявки на лицензированные данные	не проверено	открыта: нужна для заявок
GeoNetwork 4.4	ядро: пять ролей в группах, права группы на запись	ядро: черновик, проверка, одобрение; по умолчанию одобрение можно обойти	ядро: отдельное право на скачивание загруженных файлов	нет	ядро: LDAP, CAS, OpenID Connect, Keycloak, Shibboleth, EU Login	закрыта по умолчанию; можно с капчей и списком доменов
udata (data.gouv.fr)	ядро: администратор, редактор, частичный редактор	черновик — ядро; проверки куратором нет	не проверено	не проверено	ProConnect на OpenID Connect у data.gouv.fr	открыта; автоматическая пометка спама
DKAN 2.x	ядро: права Drupal	ядро: состояния «черновик», «опубликован», «скрыт», «в архиве»; по умолчанию набор публикуется сразу	не проверено	не проверено	модули Drupal, не проверено	не проверено
Huwise (OpenDataSoft)	ядро сервиса: права на набор для пользователей и групп	раздельное право публиковать; проверки куратором нет	ядро: набор виден в каталоге, содержимое — выбранным	ядро: запрос доступа на порталах с режимом маркетплейса	ядро: SAML, OpenID Connect	закрыта, открыта, открыта с одобрением; капча

Платформа	Роли	Проверка перед публикацией	Описание открыто, файл закрыт	Запрос доступа, эмбарго	Единый вход	Регистрация
Socrata (Tyler Data & Insights)	ядро сервиса: пять ролей на портал, три на набор	ядро: одобрение автоматическое или ручное	нет: закрытый набор не виден в каталоге	нет	ядро: OpenID Connect, SAML через поддержку	открыта, без роли на портале
ArcGIS Online, Hub	ядро сервиса: роли и уровни доступа к элементу	нет	частично: открытый элемент над защищённым сервисом требует входа	вступление в группу по заявке	ядро: SAML, OpenID Connect	у Hub Premium открыта для сообщества, роль публикатора по умолчанию
Hugging Face	ядро: роли в организации, группы ресурсов в платных тарифах	обсуждения и запросы на слияние	ядро: закрытый (gated) набор	ядро: автоматическое или ручное одобрение, своя анкета	в тарифах Team и Enterprise: SAML, OpenID Connect	открыта
Статический сайт, Git	права на репозиторий	запрос на слияние с обязательной проверкой	нет	нет	вход в GitHub или GitLab	—

Расширения CKAN и их риски

В ядре CKAN нет единого входа, проверки перед публикацией, доступа к отдельному файлу и защиты от спама сверх капчи. Всё это закрывают расширения, и у этой части экосистемы самое слабое сопровождение. Методика оценки расширения — в главе об экосистемах расширений, здесь то, что нашлось при сверке.

Задача	Расширение	Кто ведёт	Тесты на версиях CKAN	Что важно знать
Вход через SAML	ckanext-saml2auth	компания Keitaro	2.10, 2.11	последний выпуск — декабрь 2022 года, изменения идут в основной ветке
Вход через SAML	ckanext-saml	Link Digital (DataShades)	автоматических тестов нет	в журнале изменений — поддержка 2.12 (июнь 2026 года)
Вход через OpenID Connect	ckanext-oidc-pkce	Link Digital	2.9, 2.10, 2.11	по описанию сделано для Okta и с другими поставщиками не проверялось
Вход через LDAP	ckanext-ldap	Музей естественной истории (Лондон)	тестовые образы не обновлялись с сентября 2023 года	выпуск 4.1.3 — 28 сентября 2026 года; заявлены 2.9 и 2.10
Защита входа: блокировка перебора, второй фактор, правила паролей	ckanext-security	data.govt.nz (Новая Зеландия)	автоматических тестов нет	версия 4.2.0 от 5 октября 2026 года — для 2.11; нужен отдельный Redis; второй фактор включён по умолчанию для всех

Задача	Расширение	Кто ведёт	Тесты на версиях CKAN	Что важно знать
Открытое описание, закрытый файл, заявка на доступ	ckanext-restricted	EnviDat (Швейцария)	работающих тестов нет	заявлена совместимость до 2.11; требует ckanext-scheming и ещё двух расширений
Проверка перед публикацией	ckanext-approvalworkflow	компания Keitaro	2.10, 2.11	выпуск 2.0.0 — март 2026 года; включается для всего портала или отдельных организаций; нужна почта
Настраиваемый процесс с шагами и одобрениями	ckanext-workflow	Link Digital	автоматических тестов нет	новая версия переписана летом 2026 года и не выпущена
Запросы данных	ckanext-datarequests, ветка правительства Квинсленда	правительство Квинсленда	2.10, 2.11, 2.12	исходный репозиторий CoNWeT Lab брошен в 2019 году; капчи и очереди модерации нет, но запросы можно разрешить только участникам организаций
Капча на любых формах	ckanext-recaptcha	Link Digital	2.9	единственное найденное расширение против спама сверх капчи ядра

Из таблицы следуют три практических вывода. Расширение входа — самое рискованное при обновлении: если оно не работает на новой версии ядра, в портал не может войти ни один публикатор, поэтому вход проверяют на тестовом стенде до перехода. Проверка перед публикацией в CKAN — это чужое расширение, и прежде чем её требовать, стоит спросить, нельзя ли обойтись черновиком и ролями. Закрытый файл при открытом описании CKAN даёт хуже, чем научные платформы: если это главная задача, платформу лучше выбрать под неё.

Публикация без портала

Без портала ролей и проверки меньше, но они есть.

- **Git-репозиторий.** Изменения в данных и описаниях приходят запросом на слияние (pull request в GitHub, merge request в GitLab), и его можно сделать обязательным для проверки. В GitHub защита ветки требует одобрения перед слиянием, а файл `CODEOWNERS` назначает обязательных проверяющих для папок: например, отдел статистики проверяет всё в `statistics/`. В бесплатном тарифе GitHub это работает в открытых репозиториях. В GitLab в бесплатном тарифе одобрение необязательно и не блокирует слияние; обязательные правила одобрения и владельцы кода — в платных тарифах. История Git заодно служит журналом изменений. Подробнее — в главе о Git-репозиториях.
- **Закрытый репозиторий** скрывает и данные, и описание. Если данные закрыты, а о них нужно рассказать, описание кладут в открытый репозиторий или на сайт, а данные — туда, где можно настроить доступ.
- **Zenodo и Hugging Face** дают «описание открыто, файл закрыт» без своего портала. В Zenodo файлы записи закрывают и принимают запросы на доступ, в Hugging Face набор делают закрытым (gated) с автоматическим или ручным одобрением и своей анкетой.

- **Статический сайт** штатного «по запросу» не умеет: так публикуют открытое. Для закрытых данных сайт может честно описать набор, порядок и срок ответа на запрос и дать адрес, куда писать. В разметке schema.org тогда ставят `isAccessibleForFree` и описывают условия в `conditionsOfAccess`. Подробнее — в главе о статических сайтах.

Как проверить живой портал

Ручная проверка занимает пятнадцать минут.

1. Откройте пять наборов разных организаций. У каждого видна организация и контакт, а не только имя сайта.
2. Откройте портал в режиме без входа. Описания читаются, открытые файлы скачиваются. Если портал обещает закрытые данные, их описания тоже видны, а скачивание — нет, и рядом написано, как получить доступ и сколько ждать ответа.
3. Найдите выгрузку описаний (DCAT, OAI-PMH) и посмотрите, есть ли в ней условия доступа: `dct:accessRights` в DCAT, поле прав в DataCite.
4. Откройте страницу регистрации (`/user/register` у CKAN). Она либо закрыта, либо ведёт к приглашению или проверке.
5. Откройте список пользователей, раздел запросов данных, комментарии. Там нет рекламы и мошеннических ссылок.
6. Если у вас есть тестовая учётная запись редактора, проверьте, что видны только наборы своей организации и что новый набор не попадает в поиск, пока его не опубликовали.

Команды

Для портала на CKAN. Версия ядра и расширения, связанные со входом, доступом и модерацией:

```
PORTAL=https://data.gov.kg
curl -s "$PORTAL/api/3/action/status_show" \
  | jq '{ckan_version: .result.ckan_version,
         access: [.result.extensions[] |
select(test("saml|oidc|oauth|ldap|keycloak|sso|security|2fa|restricted|approval|workflow|data
request|captcha"))]}'
```

Открыта ли регистрация. При закрытой CKAN отвечает кодом 403, при открытой показывает форму с полем пароля:

```
curl -s -o /tmp/register.html -w "%{http_code}\n" "$PORTAL/user/register"
grep -c 'name="password1"' /tmp/register.html
```

Открыт ли список пользователей и есть ли в нём явный спам. Ответ 403 значит, что список закрыт. Слово для поиска — частое в именах спамных учётных записей:

```
curl -s "$PORTAL/api/3/action/user_list?q=gacor" | jq '.result | length'
```

Для Zenodo и InvenioRDM — сколько записей открыто, закрыто и под эмбарго:

```
for s in open restricted embargoed; do
```

```
  curl -s "https://zenodo.org/api/records?q=access.status:$s&type=dataset&size=1" \
    | jq --arg s "$s" '{status: $s, records: .hits.total}'
done
```

Для Dataverse — сколько файлов открыто, закрыто и под эмбарго:

```
curl -s "https://dataverse.harvard.edu/api/search?q=*&type=file&per_page=0&show_facets=true"
\
  | jq '.data.facets[0].fileAccess.labels'
```

Запускайте проверку медленно: по одному запросу к сайту, а не десятки параллельно.

Замер на 1125 порталах CKAN

12 октября 2026 года проверены все 1839 действующих каталогов на CKAN из реестра 1.27.0. Служебный адрес `status_show` ответил у 1136 из них, это 1125 разных серверов; остальные закрыли его, спрятали CKAN за своим сайтом или не ответили. У каждого сервера записаны версия ядра и расширения, открыта ли форма регистрации анонимному посетителю, открыт ли список пользователей и сколько в нём учётных записей со словом «gacor». Это слово из индонезийской рекламы онлайн-казино, в именах обычных пользователей оно не встречается.

Регистрация открыта у 123 серверов из 1125, то есть у 11%. Капча стоит у 29 из них.

Версия CKAN	Серверов	С открытой регистрацией
старше 2.9	206	58 (28%)
2.9	247	19 (8%)
2.10	411	16 (4%)
2.11	243	28 (12%)
2.12 и новее	18	2

У старых версий регистрация открыта по умолчанию, и почти треть установок старше 2.9 так и живёт. Обновление само по себе не помогает: на 2.11 регистрация открыта у 12% серверов, то есть её включили сознательно или перенесли настройку при обновлении. Среди них — национальный портал Кыргызстана data.gov.kg на CKAN 2.11.0.

Список пользователей открыт у 547 серверов, почти у половины. В CKAN он открыт по умолчанию (`ckan.auth.public_user_details`), даже когда регистрация закрыта, а документация советует закрывать его вместе с регистрацией.

Спам там, где открыта регистрация. Среди 547 серверов с открытым списком учётные записи со словом «gacor» нашлись у 38, всего 337 записей. Из 81 сервера, где открыты и регистрация, и список, спам есть у 31, то есть у 38%. Из 466 серверов с закрытой регистрацией — у 7: это следы того времени, когда регистрация была открыта. На портале Министерства юстиции Португалии dados.justica.gov.pt (CKAN 2.7.2) таких записей 91. На data.gov.kg — 19, с именами вроде `agenslotgacorslotgacor`, и заведены они в 2026 году. Одно слово даёт нижнюю границу: у спамеров много других имён.

Расширения входа и доступа стоят на немногих порталах. На 1125 серверах: ckanext-security — 47, вход через SAML — 27, LDAP — 23, OAuth2 — 21, OpenID Connect с PKCE — 10, Keycloak — 7, закрытые файлы ckanext-restricted — 13, запросы данных — 19, модерация — 3. Отдельно — двухфакторный вход в тайской типовой сборке на 160 серверах. Имена в ответе `status_show` — это имена модулей, а не репозиториев, поэтому одно имя может означать разные ветки одного расширения. Из 19 порталов с запросами данных у 12 открыта регистрация: это ровно то сочетание, на котором вырос спам на портале метеослужбы Туниса.

Что из замера следует. Первое, что стоит сделать владельцу портала на CKAN, — открыть свою страницу `/user/register` и список пользователей без входа. Если регистрация открыта и внешние люди не публикуют, её закрывают, а спамные учётные записи удаляют вместе с их организациями, группами и запросами. В CKAN для этого есть команда `ckan clean users`, которая удаляет учётные записи с недопустимыми файлами аватаров; остальное приходится искать по именам и содержимому.

Что из этого следует

- Начните с двух ролей — администратор и редактор организации — и черновика. Модерацию вводите, только когда публикуют люди, которым вы не доверяете выложить данные сразу, и когда есть кому её вести.
- Описание набора оставляйте открытым при любом уровне доступа к файлу. Условия доступа, порядок и срок ответа на запрос пишите на странице набора и в выгрузке: `dct:accessRights` в DCAT, право доступа по словарю COAR в DataCite, `isAccessibleForFree` в schema.org.
- Если ограниченный доступ — главная задача каталога, выбирайте платформу, где он в ядре: Dataverse, InvenioRDM, DSpace, NADA. У CKAN закрытый набор скрыт вместе с описанием, а закрытый файл — чужое расширение.
- Закройте регистрацию, если внешние люди не публикуют, и закройте список пользователей. Если публикуют, нужны приглашение или проверка, капча и ограничение прав новых учётных записей. Проверьте это на своём портале сегодня: каждый девятый портал на CKAN открыт для регистрации, и треть таких порталов уже со спамом.
- Единый вход ставьте, когда публикаторов много и их некому заводить и отключать вручную. В CKAN это расширение: проверяйте его на тестовом стенде перед каждым обновлением ядра.
- Читателю учётная запись не нужна. Описания и открытые файлы — без входа, ключа и капчи.

Источники

- re3data, схема описания репозиториев 4.0 и XSD — доступ к репозиторию, к данным и к загрузке как разные вопросы
- CKAN 2.12: организации, роли и соавторы, параметры настройки — `ckan.auth.create_user_via_web`, `ckan.auth.public_user_details`, `ckan.auth.user_create_organizations`, `ckan.auth.allow_dataset_collaborators`,

reCAPTCHA; API и токены; журнал изменений CKAN и PR 7208 — закрытая по умолчанию регистрация «чтобы не открыть сайт для спама»; значение по умолчанию сверено в коде тегов 2.9.7, 2.9.8, 2.10.3 и 2.10.4; исходный код тега ckan-2.12.0 — черновики, модуль `activity`, `user_invite`, `expire_api_token`

- Спам на порталах CKAN: закрытие регистрации на DataHub.io, май 2013 года; demo.ckan.org, заполненный спамом, 2021; команда очистки учётных записей; И. Бегтин, о цене открытости, Telegram, 25.11.2025 — портал метеослужбы Туниса
- Расширения CKAN: ckanext-saml2auth, ckanext-saml, ckanext-oidc-pkce, ckanext-ldap, ckanext-security, ckanext-restricted, ckanext-approvalworkflow, ckanext-workflow, ckanext-datarequests, ckanext-recaptcha — файлы автоматических тестов, выпуски и даты изменений, проверено 12 октября 2026 года
- Dataverse: роли и права, закрытые файлы, условия доступа, гостевая книга, эмбарго, проверка перед публикацией, способы входа
- InvenioRDM: закрытые записи и запросы на доступ, сообщества; Zenodo: доступ к записям
- DSpace (документация линейки 9.x, механизмы те же в 10.x): права на запись и файл, эмбарго, аутентификация и авторизация
- Типы доступа к данным в NADA
- GeoNetwork 4.4: пользователи и группы, права на запись, жизненный цикл записи, самостоятельная регистрация, способы входа; значения по умолчанию — data-db-default.sql
- udata: роли в организации, журнал изменений — частичный редактор в 15.4.0, пометка спама; data.gouv.fr: публикация набора и черновик, учётная запись и ProConnect
- DKAN: процесс публикации dkan_publishing, роли и права
- Huwise (OpenDataSoft): права на набор и запрос доступа, настройки безопасности портала, права в рабочем пространстве
- Socrata (Tyler Data & Insights): роли, одобрение перед публикацией, единый вход, каталог и закрытые наборы
- ArcGIS: уровни доступа к элементам, роли, защищённые сервисы, учётные записи сообщества Hub Premium
- Hugging Face: закрытые (gated) наборы, роли в организации, единый вход
- GitHub: владельцы кода, защита веток; GitLab: одобрение запросов на слияние
- Условия доступа в метаданных: DCAT 3, DCAT-AP 3.0.1, словарь видов доступа Бюро публикаций ЕС, HealthDCAT-AP, словарь прав доступа COAR, правила OpenAIRE, schema.org conditionsOfAccess, разметка наборов для Google
- Акт об управлении данными, регламент (ЕС) 2022/868 — статьи 5 и 8; европейский реестр охраняемых данных и его требования к сбору описаний; Европейское пространство медицинских данных, регламент (ЕС) 2025/327 — статьи 77 и 105
- Вход публикаторов: методические рекомендации по публикации открытых данных, версия 4.0, декабрь 2024 года; публикация описаний на data.norge.no; ProConnect
- Стандарты входа: SAML 2.0, OpenID Connect Core 1.0, OAuth 2.0, RFC 6749
- Замер 12 октября 2026 года: API поиска Zenodo и Harvard Dataverse; 1839 действующих каталогов на CKAN из реестра 1.27.0, адреса `status_show`, `/user/register` и `user_list`

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года — запросы в приложении
- Лицензии и условия доступа, Каталоги микроданных, CKAN и его расширения

Скорость развёртывания и сопровождение

Быстрый запуск и дешёвые десять лет — разные ответы, и их записывают отдельно. Каталог закрывают, когда его некому обновлять, а не когда платформа «плохая».

Запуск виден в первую неделю: появилась страница и первый набор. Сопровождение видно через три года, когда кончился проект, а каталог ещё должен открываться, принимать правки и ставить исправления безопасности.

Что именно оценивают

Две величины.

Время до первого настоящего набора. Не до слайда и не до демонстрационных данных поставщика. Набор ваш: с лицензией, файлом и описанием, которое попадает в выгрузку.

Цена следующего года. Кто обновляет версию, кто проверяет расширения, кто отвечает на письмо «ссылка не открывается», где лежат копии, что будет, если тариф или подрядчик кончатся.

Платформа с короткой установкой и дорогим сопровождением годится для пилота. Платформа, которую некому обновлять, не годится для каталога, который должен пережить грант.

Ориентиры по срокам

Сроки — порядок величины для команды, которая уже умеет администрировать серверы. Пилот с десятью наборами короче. Каталог с доработками и переносом старых данных длиннее.

Способ	До первого набора	Кто обновляет потом
Публикация в чужом каталоге	часы	свои описания и файлы; сервер — чужой
Статический сайт или репозиторий	часы или дни	файлы и страницы; сервер почти не нужен
Облачный сервис	дни	поставщик; вы платите по тарифу
Свободное ПО в готовом контейнере	дни или недели	своя команда: сервер, копии, версия
Свободное ПО с расширениями и своей схемой	недели или месяцы	то же, плюс проверка каждого расширения
Своя разработка	месяцы	всё, включая безопасность и чужие форматы

Цифры не обещание подрядчика. Они нужны, чтобы сравнить способы между собой, а не чтобы поставить срок в приказе.

Что съедает время после запуска

- **Обновления безопасности.** Платформу, которую не обновляли два года, уже нельзя просто «поднять версию»: по дороге сменились расширения, схема и способ хранения файлов.

- **Расширения.** Каждое проверяют на новой версии ядра. Одно заброшенное расширение удерживает весь портал на старой версии. Как это оценивать — в главе об экосистемах расширений.
- **Копии и место.** Своё хранилище файлов растёт незаметно. Копия, которую никто не пробовал развернуть, копией не считается.
- **Люди.** Нужен хотя бы один человек, который отвечает на письма и выпускает обновления. Без этого даже удачный запуск превращается в заброшенный сайт. Что делать, когда каталог всё-таки закрывают, — в главах о жизненном цикле и об архивации.

Быстрый запуск важнее для пилота и разовой публикации. Низкая цена сопровождения важнее для каталога, который должен прожить десять лет. Как оформить такой выбор — в главе «Как обосновать решение».

Когда что выбирать

Ситуация	Что важнее
Нужно показать десять наборов до конца квартала, команда на год не обещана	чужой каталог, статический сайт или облако
Каталог войдёт в закон или в грант на много лет, есть ставка администратора	свободная платформа, срок запуска вторичен
Обязательные функции живут только в расширениях, которые не тестируют на текущем ядре	другой способ: облако, где функция в сервисе, или другая платформа
Подрядчик поставил портал и уходит вместе с актом	в договор входит год сопровождения и выгрузка всех описаний и файлов, иначе запуск был чужим

Как это сделано в платформах

Сверено с документацией платформ в октябре 2026 года. Срок в клетке — тот же порядок величины, что в таблице способов, не обещание.

Платформа	До пилота	Как живут версии	Что проверить до установки
CKAN	дни или недели: есть официальная сборка в Docker	полностью поддерживается последняя минорная версия (с августа 2026 года — 2.12); предыдущая получает исправления ошибок и безопасности	каждое расширение на этой версии ядра; см. CKAN и расширения
InvenioRDM	первая установка тяжелее статического сайта	одна крупная версия в год; исправления — не меньше года и ещё полгода после следующей. v14 вышла 7 августа 2026 года, v13 поддерживают до 7 февраля 2027 года	успеете ли обновляться раз в год
Dataverse	есть инструкция установки и обновления	с версии 6.8 контейнеры считают готовыми для рабочих установок; несколько серверов приложений в контейнерах не поддерживаются	договор с DataCite — отдельно от установки. Текущий выпуск — 6.12.1, живые сайты отстают
DSpace	как у других научных репозиториев	официальная инструкция обновления; текущий выпуск — 10.1, в реестре ещё встречаются 8.x и 9.x	библиотеке, у которой DSpace уже стоит, обновление дешевле новой установки

Платформа	До пилота	Как живут версии	Что проверить до установки
GeoNetwork	официальная инструкция, в том числе Docker	профили страны — плагины схем, их проверяют при обновлении	рядом обычно нужен картографический сервер: GeoNetwork описывает слой, сервер его отдаёт
Huwise, Socrata, ArcGIS Hub	дни	обновления делает поставщик	как забрать все описания и файлы, если тариф кончится
Статический сайт	часы или дни	нечему падать в привычном смысле	кто соберёт сайт, когда уйдёт единственный, кто это умел

У научных платформ первая установка тяжелее статического сайта, зато DOI, протоколы и роли уже в ядре. У облака срок запуска короткий, пока устраивают тариф и место размещения. Уйти позже стоит столько, сколько стоит выгрузка и перенос адресов. Разбор цены ухода — в главе о моделях создания.

Как оценить до покупки или установки

- Есть ли официальный контейнер и короткая инструкция обновления.
- Сколько последних версий ещё поддерживаются и как часто они выходят.
- Сколько расширений нужно для обязательных функций и кто их выпускает.
- Кто будет дежурить через два года: своя ставка, подрядчик или поставщик сервиса.
- Можно ли выгрузить все описания и все файлы. Если нет, короткий запуск оплачивается дорогим уходом.

Если на вопрос о дежурстве нет ответа, более простой способ публикации обычно честнее красивого портала.

Как проверить живой портал

Версия видна снаружи не у всех платформ. Там, где видна, её сравнивают с поддерживаемой, а не с «самой новой вообще».

У CKAN версию и список плагинов отдаёт служебный адрес:

```
curl -s "https://portal.example/api/3/action/status_show"
```

В ответе смотрите `ckan_version` и `extensions`. Версия на два поколения старше поддерживаемой — повод спросить, кто и когда обновлял портал. Расширение из списка, которого нет в вашей матрице, — повод спросить, зачем оно стоит.

У облачного сервиса версии снаружи часто нет. Тогда проверяют другое: открывается ли выгрузка всех описаний и можно ли скачать файл без входа. Это и есть то, что вы заберёте с собой.

Дата в подвале «обновлено в 2019 году» при живом сайте сама по себе мало значит: подвал забывают. Дата последнего изменённого набора и возраст версии говорят больше.

Источники

- Установка CKAN в Docker, обновление CKAN, политика поддержки версий CKAN
- Политика поддержки InvenioRDM — сроки поддержки v13 и v14, проверено 11 октября 2026 года
- Установка Dataverse, обновление, контейнеры в работе
- Установка GeoNetwork, в Docker
- Экосистемы расширений
- Свои силы, свободное ПО или облачный сервис

Другие возможности

Поиск, предпросмотр, несколько языков, сбор чужих каталогов, статистика и обратная связь попадают в техническое задание, потому что они есть у большого портала. У каждой есть порог. Пока он не пройден, раньше них делают выгрузку описаний, лицензию и постоянные адреса.

Коротко:

- Поиск по тексту описаний и фильтры нужны, когда наборов сотни и люди ищут по теме. Десяткам наборов с понятными названиями хватает списка.
- Предпросмотр помогает выбрать файл. Ссылка в статье и запрос программы идут на файл и API: предпросмотр для них не источник.
- Перевод кнопок и перевод названий — разные работы. Поиск на втором языке находит набор, только если на этом языке записано название или описание.
- Сбор чужих каталогов нужен тому, кто показывает чужие наборы. Порталу со своими наборами он добавляет карточки без владельца.
- Счётчик просмотров страницы не видит скачивание через API. В годовой отчёт такая цифра не годится: как числа каталога обманывают, разобрано в главе о метриках.
- Форма без человека, который её читает и отвечает в названный срок, хуже контакта в карточке.

Как процесс сбора устроен у источника и у сборщика — в главе «Федерации и агрегация». Здесь — порог, место возможности в платформе и проверка живого сайта.

Когда какая нужна

Возможность	Нужна, когда	Можно обойтись, когда
Поиск и фильтры	наборов сотни и больше, ищут по теме, организации и формату	наборов десятки, у каждого понятное название
Предпросмотр	человек должен понять файл, не скачивая его	файлы маленькие, или данные читает только программа
Несколько языков	несколько официальных языков или международная аудитория, и есть кому вести перевод описаний	один город и один язык
Сбор чужих каталогов	региональный, национальный портал или агрегатор	портал публикует только своё
Статистика использования	по ней решают, что обновлять, и считают и страницы, и API	нужна одна «большая цифра» в годовом отчёте
Обратная связь	публичный портал, и назван человек, который читает сообщения	архив по регламенту гранта: хватает контакта в описании

Поиск через API — соседняя возможность, она разобрана в главе об API и выгрузке. На CKAN строка на сайте и запрос программы смотрят в один и тот же индекс, поэтому слово, которого нет в выдаче сайта, не найдётся и через API.

Поиск и фильтры

Поиск по словам и фильтры по организации, теме и формату. На геопортале к ним добавляется поиск по карте.

Достаточно списка, когда наборов десятки. Отдельный поисковый индекс — копия описаний, устроенная для быстрого поиска, — нужен, когда наборов тысячи и люди фильтруют, а не листают. Если индекс разошёлся с базой описаний, набор есть, а поиск его не показывает.

Поиск по названию и поиск по описанию — разные настройки индекса. Слово, которым набор ищут, часто стоит в тексте карточки, а не в заголовке. Индекс, в который описание не попало, такое слово не найдёт.

У CKAN индекс — Solr, он входит в обычную установку. Обычный запрос ищет по общему полю `text`: туда попадают название, описание, теги и другие поля карточки. Запрос с двоеточием ищет по одному полю, например `title:бюджет`. Фильтры по организации, группе, тегу и формату файла тоже в ядре. У Dataverse 6.12 Solr — обязательная часть установки, и его API нельзя открывать наружу: кто достучится до Solr, сможет менять индекс мимо каталога. У GeoNetwork 4.4 индекс — Elasticsearch (в документации октября 2026 года рекомендована версия 8.19.13, минимальная — 8.14); порт самого индекса наружу не открывают, поиск идёт через каталог. У InvenioRDM в инструкции установки назван OpenSearch. У udata поиск — Elasticsearch, и он не обновляется сам в момент записи карточки.

14 октября 2026 года на data.gov.gr (CKAN 2.11.3) запрос со словом `Guarantees` вернул 6 наборов. В греческом названии первого этого слова нет: оно стоит в английском переводе названия. Значит, перевод названия попал в индекс. Фильтр по формату CSV в тот же день отобрал 5599 наборов из каталога: фильтр по формату на этом портале работает.

Проверка: наберите слово из середины описания, а не из названия. Затем отберите одну организацию и один формат. Если фильтра нет, а наборов больше нескольких сотен, людям приходится листать выдачу.

Предпросмотр

Таблица, график или карта на странице набора помогают решить, тот ли это файл. Предпросмотр нельзя процитировать и нельзя надёжно забрать программой. Копия для показа устаревает отдельно от файла: её проверяют так же, как исходный файл. Об этом — в главе «Анатомия каталога».

У CKAN виды показа — плагины из поставки ядра: таблица (`datatables_view`), текст, картинка, видео, аудио, встроенная страница. Таблица включается, только если строки лежат в DataStore. Сам DataStore — часть ядра, а заливку CSV в него делает расширение XLoader. В документации CKAN 2.12 просмотрщиков Recline уже нет: таблицу показывает `datatables_view`. Карту слоя даёт расширение ckanext-geoview, выпуск 0.3.2 от 2 октября 2026 года. Файл на чужом сайте браузер со страницы CKAN сам не покажет: для этого включают прокси ресурсов, и у прокси есть предел размера. Встроенную страницу (`webpage_view`) не включают там, где наборы создаёт кто угодно: она показывает чужой адрес внутри портала.

На data.gov.gr 14 октября 2026 года включены и `datatables_view`, и `datastore`, и `xloader`. У набора о государственных гарантиях за декабрь 2023 года файл XLS в DataStore не загружен. CSV того же набора загружен, но поле `datastore_contains_all_records_of_source_file` равно `false`: в базе лежит не весь файл. Предпросмотр таблицы зависит от конкретного файла и может показать только его часть.

У Dataverse механизм подключения просмотрщиков есть в ядре (внешние инструменты), а сами просмотрщики — отдельный набор сообщества, dataverse-previewers. Их ставит администратор. На странице файла предпросмотр закрытого файла не показывают. Его видит тот, у кого уже есть право скачать файл, в том числе по отдельной ссылке предпросмотра черновика. Поэтому фраза «у Dataverse есть предпросмотр» верна для установки, где просмотрщики включены, и для файла, который этому человеку можно скачать.

У InvenioRDM просмотрщики входят в поставку: PDF, CSV, картинки, Markdown, JSON, XML, ZIP, блокноты Jupyter; с выпуска v12 ещё аудио и видео. Часть из них, включая CSV, читает файл из хранилища самой установки. У Socrata страница набора — это таблица и холст исследования: руководство по статистике называет их обычными страницами просмотра. У GeoNetwork в ядре карта слоя, а не таблица CSV.

Проверка: откройте набор с таблицей и набор с закрытым файлом. У первого на странице есть таблица или прямая ссылка на скачивание, без обещания таблицы, которой нет. У второго содержимое не видно тому, кому файл закрыт.

Несколько языков

Здесь три разные вещи.

Язык кнопок. Меню, подписи, сообщения об ошибках. У CKAN переводы интерфейса лежат в ядре: список языков задают настройками `ckan.locales_offered` и `ckan.locale_default`. У Dataverse и InvenioRDM дополнительные языки интерфейса включаются настройкой, переводы поставляет сообщество. У GeoNetwork 4.4 список языков интерфейса выбирают в настройках; если язык один, переключатель не показывают.

Язык данных. Поле вроде `dct:language` говорит, на каком языке сама таблица или текст. Это не перевод карточки.

Перевод карточки. Название и описание на каждом языке, на котором порталом пользуются. У CKAN такие поля делает расширение ckanext-fluent вместе с ckanext-scheming. Оба ведёт сообщество CKAN, автоматические тесты на основной ветке 14 октября 2026 года включают CKAN 2.10, 2.11 и 2.12. У GeoNetwork несколько языков в самой записи ISO — возможность ядра: название и аннотация хранятся отдельно для каждого языка, и поиск можно ограничить языком интерфейса или вести по всем языкам записи. У Dataverse и InvenioRDM отдельные переводы полей появляются, если это заложено в схеме. Само поле «язык набора» перевод названия не заменяет.

Если кнопки на двух языках, а карточки на одном, поиск на втором языке пуст: человек переключил язык и не находит набор, который лежит под названием на первом.

data.gov.gr как раз ведёт карточки на двух языках: в списке плагинов есть `fluent` и `scheming`. У набора о государственных гарантиях за декабрь 2022 года греческое название, греческое описание и английское название говорят о декабре 2022 года, а английское описание — «Hellenic Republic Guarantees – January 2026». Перевод описания указывает на другой месяц. Поиск по английскому слову при этом работает, потому что английское название в индекс попало. Перевод, который никто не сверяет с исходной карточкой, расходится с ней тихо.

Машинный перевод собранных карточек делает их находимыми на языках ЕС: data.europa.eu так переводит названия и описания уже после сбора. Точную формулировку на двух языках по-прежнему пишет издатель. Как устроена цепочка сбора — в главе о федерациях.

Проверка: переключите язык и найдите набор словом на этом языке. Затем откройте карточку и сравните перевод описания с текстом на языке оригинала, а не только с переводом названия.

Сбор чужих каталогов

Портал забирает описания из других каталогов по расписанию: из DCAT, API CKAN, CSW, OAI-PMH. Берут карточки, не файлы. Региональному и национальному порталу и любому агрегатору это основная работа. Ведомству, которое публикует только своё, сбор не нужен: чужие карточки увеличивают счётчик и оставляют вторую копию без владельца.

Где это живёт в платформе, сверено в октябре 2026 года:

- **CKAN.** Каркас сбора — расширение ckanext-harvest. Сбор RDF делает ckanext-dcat, сбор геокаталогов — ckanext-spatial. Кто их ведёт и с какой версией ядра они работают — в главе про CKAN.
- **udata.** Шесть сборщиков входят в поставку: DCAT, два варианта CSW, CKAN, DKAN и формат французского министерства сельского хозяйства. Ни один не включён, пока его нет в настройке `HARVESTER_BACKENDS`. Новый источник администратор ещё и подтверждает, иначе сбор не стартует. Запись в базу не попадает в поиск сама: после прохода индекс пересобирают отдельно. На data.gouv.fr 14 октября 2026 года список источников отдавал 502 записи. В ту же ночь источник «Communauté de Communes du Pays de Landivisiau» (сборщик CSW, каталог GeoBretagne) завершился без ошибок и забрал 4 набора.
- **Dataverse.** Клиент сбора есть в ядре и ходит по OAI-PMH, в том числе в другие установки Dataverse. Собранный набор в поиске ведёт на исходный репозиторий, править его у себя нельзя. Сервер OAI-PMH, которым отдают свои наборы наружу, по умолчанию выключен: это отдача, а не сбор.
- **Piveau.** Сбор DCAT — основная функция ядра.
- **GeoNetwork.** В ядре забирает CSW, OAI-PMH, сервисы OGC и коллекции STAC.
- **Huwise.** Сборщики входят в сервис: CKAN, Socrata, ArcGIS, ArcGIS Hub, файл `data.json`, data.gouv.fr и другие порталы Huwise. Сбор из CSW по умолчанию выключен, его включают через поддержку.
- **Socrata (Tyler Data & Insights).** Catalog Connector в панели администратора забирает файл `data.json` и публичный ArcGIS Server, плюс есть обмен между доменами самой Socrata. Наборы из `data.json` после синхронизации становятся открытыми: закрытый у

источника набор на стороне Socrata снова откроется. Это проверяют до подключения чужого каталога с непубличными записями.

- **JKAN и статический сайт.** Сбора в шаблоне нет: файл `data.json` собирает генератор из страниц. Чужие каталоги забирают отдельным скриптом в том же репозитории, как на opendata.scot.

У ArcGIS Hub в этой сверке сбор произвольных чужих каталогов не подтверждён. У InvenioRDM, DSpace и NADA клиент сбора в документации этой сверки не закрывался.

На data.gov.gr у набора, забранного сборщиком DCAT-AP, в карточке есть адрес оригинала (`https://portal.raise-science.eu/datasets/...`) и название источника сбора. Читатель отличает чужую публикацию от городской.

Проверка: в списке источников есть живой адрес, дата последнего успешного прохода и число забранных наборов. На карточке собранного набора видна ссылка на оригинал. Слово из названия находится поиском: запись в базу без пересборки индекса на сайте не видна.

Статистика использования

Просмотры страниц, скачивания файлов, вызовы API. По ним решают, какие наборы обновлять в первую очередь, и показывают, что каталогом пользуются. Одна цифра «просмотров портала» для этого не годится: она легко раздувается и не видит тех, кто забирает файл программой.

В документации CKAN 2.12 подсчёт просмотров — плагин `tracking` из поставки ядра. Его включают в список плагинов. В CKAN 2.10 и более ранних версиях тот же подсчёт включался настройкой внутри ядра. Плагин пишет просмотры страниц наборов и файлов. Сводка, которую показывает сайт, обновляется командой `ckan tracking update`; чтобы цифры попали в поиск, после неё пересобирают индекс. Поле `recent` — просмотры за 14 дней, `total` — все учтённые просмотры. Вызовов API этот счётчик не видит. На data.gov.gr плагин включён: 14 октября 2026 года у одного собранного набора сводка показывала 51 просмотр всего и 31 за последние 14 дней, у соседнего набора оба числа были нулями. Ноль значит, что страницу не открывали с тех пор, как подсчёт её видит.

Отдельный путь — расширение ckanext-googleanalytics (сообщество CKAN). Автоматические тесты на основной ветке идут на CKAN 2.10, 2.11 и 2.12. В главе про CKAN отмечено, что документация расширения называет версии до 2.11. Журналы веб-сервера остаются третьим способом, когда нужно увидеть запросы, которые страница не считает.

У Dataverse обычный счётчик скачиваний есть без дополнительной установки. Учёт по правилам Make Data Count — это отдельная программа Counter Processor, ежедневный запуск и обмен с DataCite. Make Data Count собирает просмотры, скачивания и цитирования исследовательских данных по правилам COUNTER для данных. Цитирования из этого обмена приходят только для наборов с DOI. Установка на Handle может посчитать просмотры и скачивания у себя, в концентратор DataCite их не отправляют.

У InvenioRDM с выпуска v12 статистика просмотров и скачиваний входит в поставку и следует тем же правилам COUNTER и Make Data Count. На карточке по умолчанию показывают уникальные просмотры и скачивания: повторный заход с того же источника

второй раз не прибавляют. Актуальный выпуск линии — v14.

У Socrata статистика входит в сервис. Набор данных Site Analytics считает просмотры страниц, скачивания и чтения через API, так что программный доступ в этой системе виден. У udata счётчики лежат в карточке набора. На data.gouv.fr 14 октября 2026 года у набора «Trajets réalisés en covoiturage» (реестр подтверждённых совместных поездок) API показывал 18 334 348 просмотров и 63 155 скачиваний файлов. Рядом стоящие числа отвечают на разные вопросы: страница и файл. В отчёт о спросе берут то число, которое совпадает со способом, которым данными пользуются.

Проверка: откройте статистику набора, который скачивают скриптом, и набора, который открывают в браузере. Если у первого нули, счётчик видит только страницы.

Обратная связь

Кнопка «нашли ошибку», обсуждение на карточке, форма «запросите набор», витрина примеров использования. Пример использования — приложение, статья или визуализация на основе набора, привязанные к его карточке. Для публичного портала органа власти это способ увидеть спрос. Для архива, куда данные кладут по регламенту гранта, достаточно контакта в описании.

Форма без человека, который её читает, хуже контакта в карточке: письма копятся, а пользователь считает, что запрос принят. Прежде чем ставить форму или витрину, называют, кто отвечает и за какой срок.

У udata обсуждения и примеры использования входят в ядро. На data.gouv.fr 14 октября 2026 года сводка портала показывала 17 732 обсуждения и 5996 примеров использования. У того же набора о совместных поездках было 23 обсуждения, из них 19 ещё открыты, и 5 примеров использования. Счётчик сообщений не говорит, ответили ли на них.

У CKAN витрину примеров делает расширение ckanext-showcase. Его ведёт сообщество CKAN, тесты на основной ветке идут на 2.10, 2.11 и 2.12, последний выпуск на 14 октября 2026 года — 1.8.4 от 22 июля 2025 года. На CKAN 2.12 для картинок витрины отдельно настраивают хранилище файлов. Страницы «о портале» делает ckanext-pages, тоже сообщество CKAN. Последний выпуск — v0.5.2 от 7 июня 2023 года. Поддержка CKAN 2.12 влита в основную ветку 26 июня 2026 года, выпуска после этого нет: на 2.12 расширение ставят из основной ветки. Тесты этой ветки включают 2.10, 2.11 и 2.12. При обновлении ядра это отдельная проверка, даже когда тесты зелёные: ставить нечего, кроме снимка ветки.

У Dataverse на карточке есть контакт, а у закрытого файла — запрос доступа. Это ядро. Книга отзывов при скачивании считает визит, письмо владельцу она не заменяет.

Проверка: отправьте сообщение об ошибке на один набор и посмотрите, дошло ли оно до владельца данных. Общий ящик, который никто не открывает, проверку не проходит.

Как это сделано в платформах

Пометки — ядро, расширение, нет, не проверено, как в сводной таблице. «Ядро» здесь значит «входит в поставку». Плагин CKAN, который лежит в той же поставке и только включается, записан как ядро. Сверено с документацией и живыми порталами 14 октября 2026 года.

Платформа	Поиск	Предпросмотр	Перевод карточек	Сбор	Статистика	Обратная связь
CKAN 2.12	ядро (Solr)	ядро (виды показа); таблица — при DataStore; заливка CSV — XLoader; карта — ckanext-geoview	расширение (ckanext-fluent и scheming); кнопки — ядро	расширение (harvest, dcat, spatial)	ядро (плагин tracking, просмотры страниц); ckanext-googleanalytics — расширение	витрина и страницы — расширения; контакт — поле карточки
Dataverse 6.12	ядро (Solr)	механизм в ядре, просмотрщики ставит администратор	язык интерфейса — настройка; перевод полей — если он есть в схеме	ядро (клиент OAI-PMH)	счётчик скачиваний — ядро; Make Data Count — отдельная программа	контакт и запрос доступа — ядро
InvenioRDM v14	ядро (OpenSearch)	ядро	язык интерфейса — настройка	не проверено	ядро с v12 (COUNTER, Make Data Count)	не проверено
GeoNetwork 4.4	ядро (Elasticsearch)	карта слоя — ядро	ядро (языки в записи ISO и в интерфейсе)	ядро	не проверено	не проверено
udata	ядро (Elasticsearch)	не проверено	не проверено	ядро (шесть сборщиков, по умолчанию выключены)	ядро хранит счётчики	ядро (обсуждения и примеры использования)
Huwise	ядро	не проверено	не проверено	ядро (CSW включают по запросу)	не проверено	не проверено
Socrata	ядро	ядро (таблица и холст исследования)	не проверено	ядро (Catalog Connector)	ядро (просмотры, скачивания, чтения API)	не проверено
ArcGIS Hub	ядро	не проверено	не проверено	не проверено	не проверено	не проверено
JKAN, статический сайт	фильтр по JSON в браузере, API нет	нет	нет: текст пишет автор	нет в шаблоне; скрипт пишут отдельно	нет	контакт в тексте страницы

Клетка «не проверено» значит, что в документации этой сверки ответа не было. Перед выбором платформы её закрывают по руководству этой платформы.

Пример: город копирует большой портал

Коротко, как выглядит обоснование по шаблону карточки решения:

- **Контекст.** Город публикует около 40 наборов, один официальный язык, чужих каталогов нет. В техническое задание перенесли список национального портала: поиск с фильтрами, предпросмотр таблиц и карт, три языка, сбор порталов районов, статистика для годового отчёта, форма «нашли ошибку» и витрина примеров.
- **Варианты.** Требовать весь список и закрыть его расширениями CKAN. Взять Socrata или другой облачный сервис, где таблица и статистика уже входят в сервис. Записать в требования поиск и контакт, а остальное — как отказ с порогом пересмотра.

- **Критерии.** Число наборов и число языков, на которых карточки реально будут вести. Есть ли чужие каталоги, которые город обязан забирать. Назван ли человек, который читает форму и обновляет переводы.
- **Решение.** В требованиях остаются поиск по названию и описанию и контакт в карточке. Предпросмотр таблицы желателен только вместе с хранилищем таблиц, для тех CSV, которые туда грузят. Перевод карточек, сбор, витрина и статистика для отчёта в эту очередь не входят.
- **Почему.** Весь список расширений CKAN не проходит порог: второго языка нет, районы свои каталоги городу не отдают, форму некому читать, а каждое расширение проверяют при обновлении ядра. Socrata, где таблица и статистика уже в сервисе, не проходит тот же порог: эти функции не становятся задачей от того, что они входят в тариф. Поиск и контакт закрывают сегодняшние 40 наборов.
- **Когда пересмотреть.** Наборов стало больше нескольких сотен, и слово из описания не находится. Появился второй официальный язык и назван человек, который ведёт перевод карточек. Городу поручили забирать каталоги районов. На форму назначен сотрудник и срок ответа.

Публикация без портала

На статическом сайте и в Git-репозитории эти шесть возможностей почти все остаются на авторе страницы.

Поиск — это поиск по сайту силами площадки или список с понятными названиями. Отдельный индекс автор добавляет сам, когда страниц становится много. Предпросмотра в смысле каталога нет. GitHub показывает небольшой CSV и TSV таблицей; пределы площадки разобраны в главе о форматах. Язык один, тот, на котором написана страница; второй язык — вторая страница, и её тоже нужно найти поиском. Сбора чужих каталогов нет. Статистика площадки считает визиты репозитория или сайта и пропускает тех, кто забрал файл прямой ссылкой. Обратная связь — адрес в тексте страницы и вопросы в репозитории, если их кто-то читает.

Этого хватает для небольшого набора страниц. Пороги из таблицы выше здесь те же: сотни файлов, второй язык, обязанность забирать чужие каталоги и отчёт о спросе по API возвращают к каталогу. Как сделать такую публикацию находимой — в главе «Как сделать публикацию находимой».

Как проверить живой портал

Команды ниже выполнены 14 октября 2026 года.

Поиск по слову, которого нет в названии на языке по умолчанию, но которое есть в переводе названия. Ответ CKAN — число найденных наборов в `result.count`:

```
curl -s "https://data.gov.gr/api/3/action/package_search?q=Guarantees&rows=1"
```

В тот день счётчик был 6. У первого набора греческое название — про государственные гарантии за декабрь 2023 года, английское содержит слово Guarantees.

Фильтр по формату. Счётчик в том же поле:

```
curl -s "https://data.gov.gr/api/3/action/package_search?fq=res_format:CSV&rows=0"
```

Сверка перевода с исходной карточкой. Смотрите `title_translated` и `notes_translated`:

```
curl -s "https://data.gov.gr/api/3/action/package_show?id=eggyiseis-ellinikoy-dimosioy-dekemvrioy-2022"
```

Просмотры страниц, если на CKAN включён плагин tracking. В ответе поиска или карточки — объект `tracking_summary` с полями `total` и `recent`.

Сводка портала на udata: число наборов, обсуждений и примеров использования — в `metrics`.

```
curl -s "https://www.data.gouv.fr/api/1/site/"
```

Список источников сбора и число записей — в поле `total`:

```
curl -s "https://www.data.gouv.fr/api/1/harvest/sources/?page_size=1"
```

У одного источника смотрят `last_job.status`, `last_job.ended` и `last_job.items`. У источника Landivisiau в ту ночь статус был `done`, ошибок не было, наборов — 4.

На любом другом портале те же вопросы без этих адресов: слово из середины описания находится; фильтр по организации и формату есть, когда наборов сотни; у закрытого файла предпросмотр пуст; после смены языка поиск на этом языке что-то находит, и перевод описания совпадает с оригиналом; у собранного набора есть ссылка на источник и дата последнего прохода; счётчик скачивания скриптом не равен нулю, если скрипты — главный способ доступа; сообщение об ошибке доходит до владельца данных.

Источники

- Поиск в руководстве CKAN 2.12, просмотр данных, подсчёт просмотров, DataStore, языки интерфейса — ключи `ckan.locales_offered` и `ckan.locale_default` в объявлении настроек CKAN 2.12
- ckanext-fluent, ckanext-scheming, ckanext-harvest, ckanext-spatial, ckanext-googleanalytics, ckanext-showcase, ckanext-pages, XLoader, ckanext-geoview — файлы автоматических тестов и выпуски, 14 октября 2026 года. У ckanext-pages последний выпуск v0.5.2 — 7 июня 2023 года, поддержка CKAN 2.12 влита в основную ветку 26 июня 2026 года. У ckanext-showcase последний выпуск — 1.8.4 от 22 июля 2025 года
- Внешние инструменты Dataverse 6.12, dataverse-previewers, предпросмотр и закрытые файлы, клиент сбора, Make Data Count, Solr в требованиях к установке, языки интерфейса
- Просмотрщики Invenio, статистика в примечаниях к v12, языки интерфейса, OpenSearch — инструкция установки InvenioRDM
- Языки интерфейса GeoNetwork 4.4, многоязычные поля в индексе, установка Elasticsearch

- Сбор в udata — шесть сборщиков, по умолчанию выключены, поиск обновляется отдельной пересборкой индекса
- Сборщики Huwise, проверено 14 октября 2026 года; сбор из CSW по умолчанию выключен — CSW harvester
- Catalog Connector у Socrata — наборы из `data.json` после синхронизации становятся открытыми; Site Analytics: просмотры, скачивания и чтения API
- Piveau
- API data.gov.gr и data.gouv.fr, 14 октября 2026 года: поиск по слову Guarantees, фильтр CSV, карточка набора за декабрь 2022 года, сводка портала и список источников сбора
- Федерации и агрегация
- Карта возможностей

ЧАСТЬ V.

Платформы и расширения

На каком программном обеспечении строят каталоги данных, что есть в ядре платформ и что добавляют экосистемы расширений.

1. Классы программного обеспечения
2. Экосистемы расширений
3. CKAN: ядро и расширения
4. Карточки платформ

Классы программного обеспечения

Тип каталога говорит, **что** публикуют. Класс программы говорит, **кто** её обновляет, можно ли добавить функцию и сколько стоит уйти. Внутри одного класса платформы различаются расширением: см. экосистемы расширений.

Классы

Класс	Примеры	Когда подходит
Свободная платформа портала	CKAN, DKAN, uData, Piveau	есть своя команда или подрядчик, нужна своя схема и свои расширения
Научный репозиторий	Dataverse, DSpace, InvenioRDM	нужны DOI, версии и цитирование
Геокаталог	GeoNetwork, GeoNode; облачный ArcGIS Hub	каталог слоёв и сервисов
Платформа показателей	PxWeb, .Stat Suite, OpenSDG	официальная статистика и показатели
Каталог микроданных	NADA	обследования с разными уровнями доступа
Облачный сервис	Huwise (OpenDataSoft), Socrata, Figshare	нет своей эксплуатации, есть абонентская плата
Сайт на CMS	WordPress, Drupal	каталог — раздел существующего сайта, и CMS умеет отдать описания машинам
Сервер протокола	THREDDS, ERDDAP, сервер STAC, SPARQL	данные читают программы, сайт-витрина вторичен
Картографический сервер	GeoServer, MapServer, QGIS Server	отдаёт слои; каталогом сам по себе не является
Отраслевая система	GBIF IPT и подобные сети	данные одной научной дисциплины
Своя разработка	сайт, написанный под одну организацию	ни одна платформа не закрывает обязательную функцию

Картографический сервер часто стоит **рядом** с геокаталогом, а не вместо него. GeoServer отдаёт карту, GeoNetwork описывает слои так, чтобы их нашли. Сервер протокола так же стоит рядом с каталогом: THREDDS и ERDDAP отдают массивы и таблицы программам, а карточку для человека ведёт другой сайт.

Четыре вопроса к классу

Марку сравнивают после того, как класс ответил на четыре вопроса. Ответы по маркам — в карточках и в главе о моделях создания.

Вопрос	Свободная платформа	Облачный сервис	Статический сайт или Git	Своя разработка
Кто обновляет	своя команда или подрядчик	поставщик	тот, кто правит файлы	своя команда, целиком
Можно ли добавить функцию	да, расширением или доработкой	только то, что есть в тарифе	правкой шаблона; API поиска так не появляется	да, ценой сопровождения
Что забираете при уходе	свои файлы и база, если копии делали	то, что поставщик отдаёт выгрузкой	файлы, они и так у вас	всё, что сами предусмотрели выгрузить

Вопрос	Свободная платформа	Облачный сервис	Статический сайт или Git	Своя разработка
Чего класс не делает	не появляется сам, пока нет администратора	не принимает ваш модуль	не отдаёт API поиска и не собирает чужие каталоги	не знает чужие стандарты, пока их не написали

Сайт на CMS и сервер протокола в эту таблицу не сведены, потому что их выбирают не вместо портала, а рядом с ним. CMS годится, когда каталог — раздел уже живущего сайта и движок умеет отдать описания машинам. Сервер протокола годится, когда данные читают программы, а страницы для людей вторичны.

Ошибка класса дороже ошибки марки. CKAN с расширением DOI не становится научным репозиторием: расширение ckanext-doi на CKAN 2.11 и 2.12 не тестировалось. Dataverse не становится порталом открытых данных с ежедневной федерацией DCAT-AP. ArcGIS Hub не выдаёт DOI. Эти границы записаны в карточках, и короткий список их уважает.

Что видно по живым каталогам

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года):

- самописные каталоги — около 16%. У геопорталов их 4%: там победили готовые платформы, в первую очередь ArcGIS. У каталогов показателей самописных 44%, и именно у этого типа реже всего есть API;
- экосистема ArcGIS — примерно треть геопорталов;
- платформу муниципалитета часто выбирает не он сам. Российский Gosweb — государственная платформа: городу остаётся публиковать туда, куда страна уже собирает данные. Польская e-mapa — коммерческий продукт компании Geo-System, который покупает каждая гмина, но поставщик обслуживает больше половины гмин страны, и выбор фактически сделан рынком.

Смена владельца облачной платформы или закрытие продукта бьёт сразу по сотням каталогов. Так произошло с Nesstar у каталогов микроданных: в 2015 году с версией 4.0 продукт объявили устаревшим, к 2022 году его окончательно сняли, и архивы переносили коллекции на другие платформы — например, канадский ODESI выбрал для переезда Borealis, построенный на Dataverse. Даже переименование поставщика без смены продукта, как у OpenDataSoft, ставшей Huwise в 2025 году, — повод перечитать договор. Поэтому класс «облачный сервис» в таблице сравнения всегда идёт с вопросом: как забрать данные, если тариф кончится. Разбор — в главе о моделях создания.

Как выбрать класс, а не марку

1. Если данных мало и обновляют их редко — чужой каталог или статический сайт, не платформа.
2. Если нет своей команды — облачный сервис или публикация в уже существующем каталоге.
3. Если нужны DOI и цитирование — научный репозиторий, а не портал открытых данных с доработками.

4. Если основной продукт — слои карты — геокаталог. Картографический сервер к нему добавляют, а не ставят вместо каталога.
5. Если основной продукт — показатели с разрезами — платформа показателей, а не CKAN с тысячей крошечных наборов.
6. Если есть команда и обязательные функции, которых нет в облаке, — свободная платформа и список расширений.
7. Своя разработка — только когда пункты 1–6 не закрывают обязательную функцию. Проверочные вопросы — в той же главе о моделях создания.

Дальше сравнивают конкретные платформы по карточкам и по матрице в главе «От функций к платформе». Класс, который не ответил на вопрос «кто дежурит через два года», из списка выбывает раньше, чем у него посчитают функции.

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. Запросы — в приложении.
- Карточки платформ
- Серверы протоколов и отраслевые системы: THREDDS, ERDDAP, GeoServer, GBIF IPT
- Nesstar: история продукта, карточка Nesstar Publisher в каталоге CESSDA — распространяется без поддержки, перенос ODESI с Nesstar на Borealis
- Объявление о переименовании OpenDataSoft в Huwise
- Geo-System, итоги внедрений технологии e-Gmina в 2025 году

Экосистемы расширений

Нужная функция часто лежит не в самой платформе, а в отдельном модуле. У модуля свой автор, свои выпуски и своя совместимость с версией ядра. Сравнивать платформы по списку «что умеет коробка» значит не увидеть ни возможностей, ни цены сопровождения.

Практический вывод такой. Для каждой обязательной функции записывают, где она живёт: в поставке платформы, в названном расширении или её нет. Расширение, без которого функция не работает, проверяют до короткого списка — по репозиторию, тестам и веткам, а не по рекламному описанию. Если единственный путь к обязательной функции — модуль, который на вашей версии ядра не проверяют, платформа в короткий список не входит. Цена принятого расширения — проверка при каждом обновлении ядра.

Полная таблица расширений CKAN — в главе «CKAN: ядро и расширения». Здесь — как устроены экосистемы вообще и как читать одну из них до установки.

Что считать расширением

Под одним словом «плагин» оказываются пять разных вещей. От вида зависит, кого спрашивать при обновлении.

Настройка поставки. Файл или переключатель, который идёт вместе с платформой. Дополнительные поля InvenioRDM задаются в конфигурации установки. Схема набора DKAN — файл JSON Schema в самом DKAN. Автор тот же, что у платформы, и обновление одно.

Модуль из той же поставки, который включают. У CKAN хранилище таблиц, просмотр текста и статистика — плагины: их имена пишут в настройку `ckan.plugins`. Выпускают их вместе с ядром, отдельного репозитория у них нет. У GeoNetwork схемы ISO 19139, ISO 19115-3, Dublin Core и каталог объектов ISO 19110 входят в выпуск. Это ядро. Чужого сопровождающего у них нет.

Отдельное расширение. Свой репозиторий, свой выпуск, своя таблица версий ядра. Так устроены пакеты ckanext-* и плагины схем GeoNetwork под страну. Сопровождать их можете вы, даже если автор — сообщество платформы.

Внешняя программа. Каталог её код не содержит, а открывает по адресу. У Dataverse просмотрщик файла — отдельное приложение: в установку кладут манифест JSON с адресом, а само приложение может работать на чужом сайте. Обновление ядра его не затирает. Зато меняется договорённость, как приложение получает файл.

Облачный сервис. Свой модуль на сервер ArcGIS Hub, Huwise или Socrata поставить нельзя. Остаются виджет на странице и API. Обновления делает поставщик. Доработать сервис под свой профиль метаданных нельзя: если нужного профиля нет, это другая платформа или своя программа рядом.

Документация InvenioRDM прямо выстраивает доработки от дешёвой к дорогой. Сначала настройка. Потом файлы в ожидаемых местах: свой шаблон, своя картинка. Потом готовый модуль. Потом свой пакет в папке установки. Потом правка ядра с отдачей правки обратно в

проект. В конце — своя копия чужого репозитория, форк: вы ведёте её сами. Логотип меняют настройкой. Форк ради логотипа означает, что каждое обновление оригинала вы переносите руками.

Чем экосистемы отличаются

Платформа	Как добавляют возможность	Кто обычно ведёт	Что проверять при обновлении
CKAN	пакет Python; имя плагина включают в `ckan.plugins`	сообщество CKAN или отдельная организация	ваша версия ядра есть в автоматических тестах пакета
DKAN	модуль Drupal; сам DKAN — тоже модуль Drupal	CivicActions ведёт DKAN, остальное — модули Drupal	две линейки: Drupal и DKAN. Выпуск DKAN 4.1.5 от 7 октября 2026 года заявлен для Drupal 10 начиная с 10.2 и для Drupal 11
Dataverse	внешняя программа по манифесту или JAR-файл экспортёра формата	консорциум GDCC и отдельные проекты	манифест и способ доступа к файлу. С версии 6.1 просмотрщикам рекомендуют короткую подписанную ссылку
GeoNetwork	плагин схемы: каталог схем XML и преобразований, иногда отдельный JAR	национальные сообщества и компании	ветка плагина названа вашей линейкой GeoNetwork
InvenioRDM	настройка, файлы в ожидаемых местах, затем свой пакет `site`	вы сами, внутри установки	свои шаблоны сверить с шаблонами новой версии
DSpace	свой код в модуле `additions` поверх классов ядра	вы сами	чужого каталога плагинов нет; свою правку заново накладывают на новую версию
GeoServer	официальное расширение или модуль сообщества	OSGeo; модуль сообщества поставляют «как есть»	модуль сообщества собирают из метки той же версии. Сборка из ночного снимка с официальным выпуском часто не стыкуется
ArcGIS Hub, Huwise, Socrata	свой модуль поставить нельзя	поставщик	полная выгрузка описаний и файлов на случай ухода

У DKAN два календаря обновлений. Дыра в Drupal и дыра в DKAN закрываются разными выпусками. Страница модуля на Drupal.org обновлена 8 октября 2026 года: модуль сообщают около 164 сайтов. Это отчёт самих установок Drupal, а не подсчёт реестра каталогов. Выпуск 4.1.5 заявлен для Drupal 10 начиная с 10.2 и для Drupal 11.

У GeoServer модуль STAC лежит в поставке как ожидающий модуль сообщества: в выпуске он есть, официальной поддержки нет. Подробнее о роли GeoServer — в главе о геопорталах.

У Dataverse с версии 5.14 чужой формат выгрузки подключают JAR-файлом внутри установки. Это уже модуль, который проверяют при обновлении ядра, в отличие от просмотрщика на чужом сайте.

Как оценить расширение до установки

Идут по странице репозитория. Рекламное описание и фраза «у соседей работало» в зачёт не идут.

Свежий выпуск сам по себе совместимости не доказывает. ckanext-doi выпуска 4.0.4 вышел 22 сентября 2025 года, а тестовые образы в репозитории не менялись с 20 сентября 2023 года: в инструкции по-прежнему названы CKAN 2.9 и 2.10. Отсутствие выпуска тоже не приговор. В таблице главы про CKAN у ckanext-pages выпусков нет с 2023 года, ставят из основной ветки,

и тесты идут на CKAN 2.10, 2.11 и 2.12. Сильный признак — автоматические тесты именно на вашей версии ядра. Имя ветки `4.4.x` у плагина GeoNetwork — тот же признак, если такая ветка в репозитории есть.

Вопрос	Где смотреть	Что считается ответом
Кто выпускает	организация в адресе репозитория, файл сопровождающих	команда платформы или организация, которая отвечает на ошибки. Один человек без ответа два года и без тестов — повод искать другой модуль
С какой версией ядра работает	таблица совместимости, файл автоматических тестов, имя ветки	ваша версия названа там. Выпуск новее ядра, в котором тесты старше вашей версии, ответом не считается
Живое ли оно	дата последнего исправления и ответы на ошибки про новую версию ядра	тесты гоняются на текущей линейке, даже если отдельного выпуска давно не было
Как ставить	документация в том же репозитории	команда установки указывает на этот репозиторий и на ветку, которая в нём есть
Какая лицензия	файл лицензии	лицензия есть и разрешает вашу установку
Кто ещё это поставил	каталог экосистемы, ответ API живого портала	модуль стоит на порталах вашей версии ядра

Каталог экосистемы CKAN в октябре 2026 года знает около 1,4 тысячи расширений и 1,3 тысячи порталов с версиями ядра. Это лучшее доступное приближение, а не полная перепись: число порталов выросло, когда туда загрузили записи реестра. Надёжнее спросить сам портал, как в разделе ниже.

Два расширения одной платформы

ckanext-dcat ведёт сообщество CKAN. В документации, сверенной 11 октября 2026 года, названы версии ядра 2.10 и 2.11. Файл тестов на основной ветке 13 октября 2026 года гоняет 2.10, 2.11 и 2.12. Последний выпуск — 2.4.4 от 11 августа 2026 года. Для портала, который должны забирать поисковики и национальный каталог, это обычная часть взрослой сборки.

ckanext-doi ведёт Музей естественной истории в Лондоне. Значок в репозитории называет CKAN 2.9.9 и 2.10.1, раздел тестов — линейки 2.9 и 2.10. Последний выпуск, 4.0.4, вышел 22 сентября 2025 года. Каталог тестовых образов с тех пор не менялся: последний коммит там датирован 20 сентября 2023 года, это как раз добавление тестов на 2.10. На CKAN 2.11 и 2.12 расширение не проверялось. Если DOI обязателен, смотрят платформу, где он в ядре: Dataverse или InvenioRDM. Разбор требования — в главе об идентификаторах.

Остальные пакеты CKAN сведены в таблице главы про CKAN: колонка документации и колонка тестов расходятся ещё у нескольких строк.

Форк

Форк — копия репозитория, которую ведёт другая команда. Копии расходятся. Ставить нужно ту, чьи тесты идут на вашей версии ядра, и записать в карточке решения её адрес.

Так разошлись две копии проверки таблиц. Исходный ckanext-validation в документации заявлен только для CKAN 2.9. Живая копия — форк правительства Квинсленда: тесты на 2.10, 2.11 и 2.12. Сопровождает её правительство Квинсленда. Если поправить модуль у себя и не отдать правку обратно, появляется третья копия, и её сопровождающий — вы.

Две копии одного плагина схемы

Тот же приём на GeoNetwork. Британский профиль GEMINI 2.3 лежит в двух репозиториях.

У Astun Technology есть ветка `4.4.x`. 6 октября 2026 года в неё попало обновление инструкции. Инструкция называет GeoNetwork 4.4 рекомендуемой версией, а файл плагина — сборкой под 4.4.12. Это та копия, которую ставят на текущую линейку GeoNetwork.

В общем собрании metadata101 ветки `4.4.x` нет. Последний коммит основной ветки — 14 февраля 2023 года. Ветка `4.2.x` остановилась 27 июня 2024 года на коммите, который сам помечен как незаконченная работа.

Ловушка в инструкции Astun: команда автоматической установки указывает не на репозиторий Astun, а на metadata101 и на ветку `4.4.x`, которой там нет. Заголовок и команда под ним расходятся. Берут репозиторий и ветку, которые открыли и в которых есть коммиты под вашу версию. Команду из текста перепроверяют.

Рядом пример, где копия одна. Канадский профиль HNAP в metadata101 выпускается номером версии GeoNetwork: выпуск 4.4.12 вышел 8 июля 2026 года. Инструкция говорит ставить его на GeoNetwork 4.4 и прямо пишет, что на более ранних версиях он не проверялся.

Плохой плагин схемы ломает весь каталог, об этом предупреждает руководство администратора GeoNetwork. Его проверяют на тестовой установке до рабочего каталога.

Внешняя программа

Просмотрщики Dataverse устроены иначе. Их разрабатывал Qualitative Data Repository, сейчас собрание ведёт GDCC: версия 1.5, 19 просмотрщиков, больше 40 типов файлов. В установку Dataverse их код не кладут. Администратор регистрирует манифест, а страница файла открывает приложение по адресу. Приложение можно оставить на сайте github.io.

Риск здесь другой. Обновление Dataverse не упирается в забытый пакет внутри сервера. Зато приложение получает доступ к файлу. Документация с версии 6.1 рекомендует короткую подписанную ссылку: она действует ограниченное время и разрешает только перечисленные вызовы API. Долгоживущий ключ API документация не рекомендует после Dataverse 6.0: ключ действует как пароль пользователя и позволяет всё, что позволено этому пользователю.

Перед включением просмотрщика записывают, какой из двух способов доступа выбран.

Чем расширение опасно потом

- Его перестали выпускать, а новая версия ядра без него не ставится. Портал остаётся без исправлений безопасности. У CKAN полностью поддерживается последняя минорная версия, предыдущая получает только исправления ошибок и безопасности. Когда выйдет следующая минорная, нынешняя 2.11 потеряет и их. Одно такое расширение удерживает весь портал.
- Два расширения правят одну и ту же форму или один шаблон. По отдельности оба ставятся, вместе ломают страницу набора. Документация CKAN просит начинать имена шаблонов, настроек и функций API с имени расширения, чтобы они не столкнулись.

- Правку оставили у себя. Дальше каждое обновление оригинала переносят руками. Это уже ваша программа, со своим выпуском и своими тестами.
- Два календаря, как у DKAN и Drupal. Закрыли один и пропустили другой.
- Плагин схемы скопирован внутрь приложения. При обновлении GeoNetwork берут ветку плагина той же линейки, иначе записи этого профиля перестают открываться.
- Внешней программе выдали долгоживущий ключ. Ключ переживёт и сотрудника, и обновление каталога.
- Модуль сообщества GeoServer взят из ночной сборки и положен на официальный выпуск. Документация пишет, что так он часто не запускается: собирать нужно из метки той же версии.

Про срок сопровождения — в главе о скорости развёртывания.

Сколько расширений уже много

Считают не «сколько плагинов модно», а сколько отдельных пакетов придётся проверять и сколько из них ведёте вы сами. Пять проверенных пакетов сообщества — обычная национальная сборка, она разобрана в главе про CKAN. Один непроверенный пакет на обязательной функции дороже десятка проверенных.

Живой портал почти всегда длиннее короткого списка из карточки. 13 октября 2026 года адрес `status_show` ответил так:

Портал	Версия CKAN	Имён в списке	Что в списке видно
data.gov.gr	2.11.3	66	знакомые пакеты сообщества (dcat, scheming, harvest, fluent, showcase, pages, archiver) и свои сборщики ведомств, в том числе Аттики и Банка Греции
data.humdata.org	2.11.4	34	18 имён начинаются с `hdx_`: это код самого портала, а не каталог сообщества

Один пакет регистрирует несколько имён. У греческого портала `scheming_datasets` и `scheming_organizations` — один ckanext-scheming. Число в ответе API — верхняя оценка, а не число репозиториев. Имена из поставки CKAN (`datastore`, `datatables_view`, `activity`, `text_view`, `stats`) из этой оценки вычитают: их обновляет команда ядра.

Имена с префиксом портала вычитают в другую сторону. Их не бросит посторонний автор, потому что автора снаружи нет. Их сопровождающий — команда портала. Для HDX таких имён 18 из 34. Это нормальное состояние большого портала и одновременно статья бюджета: при переходе на следующую версию ядра эти модули проверяет тот, кто их написал.

Тот же адрес у data.gov.au 13 октября 2026 года ответил кодом 403. Список снаружи закрыт, его берёт администратор установки.

Как это записать в выборе

В таблице платформ у каждой обязательной функции три клетки: в ядре, через какое расширение и кем оно сопровождается, либо своя доработка. В карточке решения пишут адрес репозитория, ветку, версию ядра, на которой шли тесты, и что делать, если

сопровождение прекратится: остаться на старом ядре, взять названный форк или сменить платформу. Шаблон карточки — в главе «Как обосновать решение».

Решение пересматривают, когда выходит новая минорная версия ядра, когда тесты расширения перестают включать вашу версию, когда на ошибки про новую версию перестают отвечать или когда второе расширение начинает править те же шаблоны.

Платформа, у которой обязательная функция достижима только брошенным расширением, в короткий список не входит. Как из функций получается этот список — в главе о выборе функций.

Как посмотреть, что стоит на живом портале

У CKAN список включённых плагинов и версия ядра открыты по адресу `/api/3/action/status_show`. Имена в поле `extensions` — плагины, не пакеты: один пакет может дать несколько имён.

```
curl -s "https://data.gov.gr/api/3/action/status_show" \
  | jq '{version: .result.ckan_version, n: (.result.extensions | length), extensions:
.result.extensions}'
```

Дальше каждое имя, которого нет в поставке CKAN, ищут в каталоге экосистемы или в репозитории портала и прогоняют по таблице из раздела «Как оценить расширение до установки».

У GeoNetwork полный список схем виден администратору: «Метаданные и шаблоны», затем «Стандарты». Снаружи видна схема конкретной записи, а не набор установленных плагинов. У Dataverse кнопки просмотра файла видны на странице файла; манифест с адресом приложения и способом доступа смотрит администратор.

Сводка «ядро, расширение или нет» по платформам — в карточках и в приложении.

Источники

- Как писать расширения CKAN 2.12, практика имён и интерфейсов — расширение как пакет Python, плагины включаются в `ckan.plugins`
- Политика поддержки версий CKAN
- ckanext-dcat, файл тестов test.yml — версии 2.10, 2.11 и 2.12, проверено 13 октября 2026 года; выпуск 2.4.4 от 11 августа 2026 года
- ckanext-doi — выпуск 4.0.4 от 22 сентября 2025 года; значок совместимости 2.9.9 и 2.10.1; последний коммит каталога `docker/` от 20 сентября 2023 года
- ckanext-validation и форк qld-gov-au — расхождение копий, сверка в главе о CKAN 12 октября 2026 года
- Каталог экосистемы CKAN — около 1,4 тысячи расширений и 1,3 тысячи порталов, октябрь 2026 года
- Ответы `/api/3/action/status_show` порталов data.gov.gr (CKAN 2.11.3, 66 имён) и data.humdata.org (CKAN 2.11.4, 34 имени), 13 октября 2026 года. Тот же адрес data.gov.au в этот день ответил кодом 403

- Порядок доработок InvenioRDM, свой код в папке site
- Внешние инструменты Dataverse 6.12, просмотрщики GDCC — подписанные ссылки с версии 6.1, ключ API не рекомендуют после 6.0
- Экспортёры метаданных Dataverse — чужой формат как JAR с версии 5.14
- Модуль DKAN на Drupal.org — выпуск 4.1.5 от 7 октября 2026 года, Drupal 10.2 и новее или Drupal 11; страница обновлена 8 октября 2026 года
- Плагины схем GeoNetwork, предупреждение о поломке каталога
- GEMINI 2.3, ветка 4.4.x Astun Technology — коммит инструкции от 6 октября 2026 года; копия в metadata101 — ветки `4.4.x` нет, основной ветке последний коммит 14 февраля 2023 года, ветке `4.2.x` — 27 июня 2024 года
- HNAP для GeoNetwork — выпуск 4.4.12 от 8 июля 2026 года
- Модули сообщества GeoServer, OpenSearch для ДЗЗ и STAC — в поставке, официально не поддерживаются
- Модуль additions в DSpace 10

CKAN: ядро и расширения

CKAN — свободная платформа порталов открытых данных (лицензия AGPL-3.0, ассоциация CKAN). По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) на ней работает около 1 900 порталов открытых данных. Почти ни один из них не работает на голом ядре: к ядру ставят расширения под свои задачи.

Текущая версия ядра — CKAN 2.12, вышла 26 августа 2026 года. По правилам проекта полностью поддерживается последняя минорная версия, а предыдущая (сейчас 2.11) получает только исправления ошибок и безопасности. В карточке CKAN в реестре пока отмечена версия 2.11.4. Сведения о расширениях ниже сверены 11 октября 2026 года по их документации и по автоматическим тестам в репозиториях.

Что есть в ядре

Этого хватает, чтобы опубликовать наборы, раздать права и отдать данные программе:

- наборы, файлы или ссылки, организации, группы, теги;
- Action API для поиска и записи;
- загрузка файлов (FileStore) и база для таблиц с API (DataStore);
- роли и закрытые наборы;
- поиск на Solr.

В ядре нет выгрузки DCAT, разметки schema.org, своих полей сверх стандартной схемы, сбора чужих каталогов, геопоиска и DOI. Всё перечисленное — расширения.

Открытая регистрация. Параметр `ckan.auth.create_user_via_web` разрешает любому посетителю завести учётную запись. До CKAN 2.9.7 он был включён по умолчанию. В 2.9.8 его выключили, в 2.10.0–2.10.3 он снова включён, а окончательно выключен с 2.10.4 и 2.11. Порталы, которые обновлялись со старых версий, часто сохранили прежнюю настройку. На таких сайтах роботы регистрируют сотни тысяч учётных записей и публикуют от их имени рекламу и мошеннические страницы: у портала метеослужбы Туниса в 2025 году было больше 1,3 миллиона учётных записей и десятки тысяч спам-текстов в разделе запросов данных. Если публикуют только сотрудники, регистрацию выключают, а внешних публикаторов приглашает администратор организации.

Что изменилось в 2.12. Файлы стали отдельными объектами: их можно загружать и вести отдельно от ресурсов набора и держать в нескольких хранилищах сразу. У DataStore появились быстрая постраничная выдача больших таблиц и сложные фильтры. Есть и новое оформление интерфейса, которое станет основным в CKAN 3.0. Для владельца портала это значит: переход на 2.12 требует обновить базу данных и проверить расширения, которые работают с файлами. Например, ckanext-showcase на 2.12 требует отдельно настроить хранилище. В поставке 2.12 только хранение на диске: портал, который держит файлы в облаке, при переходе меняет расширение хранения. Подробнее — в главе о хранилище.

Расширения, которые закрывают типичные требования

Колонка «Тесты» показывает, на каких версиях ядра расширение проверяется автоматически при каждом изменении. Это сильный признак совместимости, даже если текст документации ещё не обновили. Колонка «Документация» — что расширение заявляет само.

Задача	Расширение	Кто ведёт	Документация	Тесты
DCAT, DCAT-AP, DCAT-US, schema.org, Croissant	ckanext-dcat	сообщество CKAN	2.10 и 2.11	2.10, 2.11, 2.12
Свои поля и формы	ckanext-scheming	сообщество CKAN	—	2.10, 2.11, 2.12
Описания на нескольких языках	ckanext-fluent	сообщество CKAN	ставится вместе со scheming	2.10, 2.11, 2.12
Забрать наборы из других каталогов	ckanext-harvest	сообщество CKAN	—	2.10, 2.11, 2.12; последний выпуск (1.6.2) старше CKAN 2.12
Геопоиск и сбор CSW	ckanext-spatial	сообщество CKAN	2.10 и 2.11	2.10, 2.11, 2.12
Залить CSV в DataStore	XLoader	сообщество CKAN	2.10, 2.11, 2.12	2.10, 2.11, 2.12
Хранить файлы в S3 и других облаках	ckanext-file-keeper-cloud — для 2.12; до 2.12 — ckanext-s3filestore	сообщество CKAN; s3filestore — компания Keitaro	file-keeper-cloud назван в документации CKAN 2.12	file-keeper-cloud: только 2.12, выпусков нет; s3filestore: 2.10 и 2.11
Проверять ссылки и хранить копии файлов	ckanext-archiver	сообщество CKAN	—	2.10, 2.11, 2.12
Проверить таблицу по словарю полей (Table Schema)	ckanext-validation	Frictionless Data; живая ветка — форк правительства Квинсленда	исходный репозиторий: только 2.9, для 2.10 и 2.11 — «нет»	исходный: 2.9; форк: 2.10, 2.11, 2.12
Страницы «о портале»	ckanext-pages	сообщество CKAN	—	2.10, 2.11, 2.12 на основной ветке; последний выпуск v0.5.2 — 7 июня 2023 года, поддержка 2.12 влита 26 июня 2026 года без нового выпуска
Примеры использования наборов	ckanext-showcase	сообщество CKAN	2.10–2.12	2.10, 2.11, 2.12
Статистика через Google Analytics	ckanext-googleanalytics	сообщество CKAN	до 2.11	2.10, 2.11, 2.12
Вход через SAML	ckanext-saml2auth	компания Keitaro	2.10 и 2.11	2.10, 2.11; последний выпуск — 2022 год
DOI в DataCite	ckanext-doi	Музей естественной истории (Лондон)	2.9 и 2.10	тестовые образы не обновлялись с сентября 2023 года, до выхода 2.11; последний выпуск — сентябрь 2025 года

Сбор RDF из чужого DCAT делает не harvest сам по себе, а сборщик внутри ckanext-dcat. Harvest даёт очередь, расписание и экран ошибок.

Три рабочие сборки

Небольшой портал, который должны находить поисковики и национальный каталог. Ядро, ckanext-dcat (профиль DCAT-AP или национальный), ckanext-scheming, если не хватает стандартных полей. XLoader — только если таблицы нужно смотреть и запрашивать по API, а не просто скачивать.

Региональный портал, который собирает муниципалитеты. Та же сборка плюс ckanext-harvest. Муниципалитетам тогда нужен не «такой же красивый сайт», а стабильная выгрузка DCAT.

Портал, где много карт. Добавляется ckanext-spatial. Если карты — главный продукт, а таблицы случайны, честнее смотреть геоплатформу, а не наращивать CKAN. Сравнение типов — в главе о гибридах.

Все три сборки на октябрь 2026 года можно ставить на CKAN 2.12: расширения из них тестируются на этой версии.

DOI в эти сборки не входит. Для портала открытых данных он обычно не нужен, а расширение ckanext-doi на CKAN 2.11 и 2.12 не тестировалось. Научным данным место в Dataverse или InvenioRDM, где DOI есть в ядре.

Цена такой сборки

Обновление CKAN — это обновление ядра и проверка каждого расширения из таблицы. Расширение сообщества CKAN, которое тестируется на новой версии ядра, предсказуемо. Расширение, у которого про новую версию молчат и документация, и тесты, — риск: либо остаётесь на старой версии, либо берёте проверку на себя.

Сейчас такой риск у двух расширений из таблицы. ckanext-doi на 2.11 и 2.12 не тестировался вовсе: его тестовые образы старше CKAN 2.11. ckanext-saml2auth не тестируется на 2.12. Портал, которому нужен вход через SAML, перед переходом на 2.12 проверяет вход на тестовом стенде или ищет другое расширение для единого входа.

Затягивать с обновлением тоже дорого. Проект называет 2.12 последним выпуском линейки 2.x, дальше планируется CKAN 3.0. Пока в поддерживаемой паре предыдущая версия — 2.11, она получает исправления ошибок и безопасности. Когда эта пара сдвинется, 2.11 их лишится. Через три пропущенных выпуска обновление превращается в перевнедрение.

Как принимать это решение, записано в главе об экосистемах и в примере карточки про обновление ядра.

Где смотреть, кто ещё ставит расширение и на какой версии ядра. Сообщество ведёт каталог экосистемы CKAN, пока в статусе бета: в октябре 2026 года в нём около 1,4 тысячи расширений и 1,3 тысячи порталов с версиями ядра. Старый адрес extensions.ckan.org перенаправляет туда же. Число порталов в каталоге выросло вдвое осенью 2026 года, когда

туда загрузили записи реестра dataportals-registry, так что это не полная перепись, а лучшее доступное приближение. Надёжнее всего — спросить сам портал: адрес `/api/3/action/status_show` отдаёт версию ядра и список включённых расширений.

Источники

- Документация CKAN 2.12, политика поддержки версий, список изменений 2.12.0, объявление о выпуске 2.12 — 2.12 назван последним выпуском линейки 2.x
- Выпуски CKAN на GitHub — 2.12.0, 2.11.6 и 2.10.11 от 26 августа 2026 года
- ckanext-dcat, ckanext-scheming, ckanext-fluent, ckanext-harvest, ckanext-spatial, XLoader, ckanext-pages, ckanext-showcase, ckanext-googleanalytics, ckanext-saml2auth, ckanext-doi — документация и файлы автоматических тестов, проверено 11 октября 2026 года; у ckanext-pages выпуск v0.5.2 от 7 июня 2023 года и слияние поддержки CKAN 2.12 от 26 июня 2026 года сверены 14 октября 2026 года
- ckanext-validation и форк qld-gov-au — таблица совместимости и файлы автоматических тестов, проверено 12 октября 2026 года
- Каталог экосистемы CKAN — расширения и порталы, проверено 12 октября 2026 года; И. Бегтин, о каталоге экосистемы, Telegram, 05.02.2026, и о загрузке в него данных реестра, 09.10.2026
- Параметр `ckan.auth.create_user_via_web` в документации 2.9 (по умолчанию включён) и 2.12 (выключен); И. Бегтин, о цене открытости, Telegram, 25.11.2025 — спам на открытых CKAN-порталах
- Число порталов и версия в карточке платформы — запрос к выгрузке реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года.

Карточки платформ

Карточка нужна, чтобы вычеркнуть платформу до закупки. Обязательная функция должна жить в самой платформе или в расширении, которое тестируют на вашей версии ядра. Если функции нет или расширение на этой версии не проверяют, платформу в короткий список не берут, даже если остальные строки таблицы совпали.

Дальше сравнивают две–четыре карточки одного класса. Возможности разобраны в части IV, полная матрица клеток — в приложении, проверка модуля — в главе об экосистемах. Число установок — выгрузка реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года). Выпуски программ сверены с сайтами проектов в октябре 2026 года. Перед закупкой номер проверяют заново.

Как читать карточку

У одной платформы обычно три номера: выпуск на сайте проекта, версия в карточке реестра и версия на живом сайте. Они расходятся. Ставят выпуск с сайта проекта и проверяют расширения именно на нём. Отставание крупной установки на одну-две версии нормально, если обновление запланировано. Отставание на несколько лет — повод спросить, кто будет его закрывать.

Платформа	Выпуск на сайте проекта	Карточка реестра 1.27.0	Живая установка
CKAN	2.12, 26 августа 2026	2.11.4	data.gov.au — 2.11.6; data.humdata.org — 2.11.4; data.gov.gr — 2.11.3
Dataverse	6.12.1, 8 октября 2026	6.9	Harvard Dataverse — 6.10.1
DSpace	10.1, 14 сентября 2026; в тот же день вышли 9.4 и 8.5	9.2	DSpace@MIT — 8.2
GeoNetwork	4.4.12, 8 июля 2026	4.4.9	каталог Ifremer Sextant 13 октября 2026 года показывал 4.4.13
NADA	5.7.1, 4 октября 2026	5.5	версия библиотеки Всемирного банка в этой сверке не снималась
udata	17.13.0, 13 октября 2026	не приведена	data.gouv.fr в тот же день отвечал как 17.12.1

Клетка «не приведена» значит, что версию из карточки программного обеспечения реестра в текст не брали. Номер для установки берут на сайте проекта.

Клетка в таблице возможностей значит одно из четырёх.

- **Ядро** — функция выходит вместе с платформой. Включать её всё равно могут часами: DOI в GeoNetwork и Dataverse требует договора с DataCite.
- **Расширение** — отдельный модуль со своим автором и своей таблицей версий ядра. Кто его ведёт, написано в карточке.
- **Нет** — эту функцию закрывают другой платформой или своей программой рядом.
- **Не проверено** — в октябре 2026 года документация ответа не дала. Перед выбором смотрят сами.

В карточке одни и те же разделы: класс и лицензия, кому она подходит, таблица возможностей, экосистема, запуск и сопровождение, ограничения, примеры.

С какой карточки начать

Строка задаёт класс. Платформу из другой строки берут вторым каталогом, когда у организации две задачи сразу. Как устроить такую пару — в главе о гибридах.

Задача	Карточки
Портал открытых данных, есть своя команда или подрядчик	CKAN, DKAN, udata, Piveau
Тот же портал, своей эксплуатации нет	Huwise и Socrata, ArcGIS Hub
Десятки наборов, их правят файлами в Git	JKAN
DOI, версии и цитирование	Dataverse, InvenioRDM; DSpace — если библиотека уже на нём; без своей установки — Figshare
Слои и сервисы карт	GeoNetwork, GeoNode, ArcGIS Hub
Показатели с разрезами	PxWeb; обмен по SDMX — .Stat Suite
Обследования с разными уровнями доступа	NADA; университетский архив с DOI — Dataverse

CKAN

Свободная платформа порталов открытых данных, лицензия AGPL-3.0, ассоциация CKAN. Текущий выпуск — 2.12 от 26 августа 2026 года, последний в линейке 2.x. Полностью поддерживается он. Предыдущий, 2.11, получает исправления ошибок и безопасности, пока пара поддерживаемых версий не сдвинется. Порталов открытых данных на CKAN в реестре — 1 906.

Кому подходит. Органу власти или организации со своей командой либо подрядчиком, если каталог должен отдавать DCAT и жить много лет.

Кому не подходит. Каталогу, которому нужен DOI в сборке, проверенной на текущем ядре: его закрывают Dataverse или InvenioRDM. Организации, у которой некому ставить обновления.

Возможность	Где
API поиска, файлы, роли, API к таблицам	в ядре
DCAT и schema.org	расширение ckanext-dcat; поддержка заявлена для 2.10 и 2.11, тесты идут и на 2.12
Свои поля	расширение ckanext-scheming; тесты на 2.10, 2.11 и 2.12
Сбор чужих каталогов	расширение ckanext-harvest; тесты на 2.10–2.12, последний выпуск 1.6.2 старше 2.12
Облачное хранилище файлов	в поставке 2.12 только диск; облако — ckanext-file-keeper-cloud, тесты на 2.12, выпусков нет
DOI	расширение ckanext-doi; на 2.11 и 2.12 не тестировалось
Цитирование	расширение: ckanext-doi показывает ссылку только при DOI, ckanext-citeproc оформляет её по стилям CSL (заявлены 2.10 и 2.11)
Единый вход через SAML	расширение ckanext-saml2auth; на 2.12 не тестируется

Возможность	Где
Карта сайта	в ядре нет

Экосистема. Большая, значимую часть ведёт само сообщество CKAN. Взрослая сборка — ядро, ckanext-dcat и, если не хватает стандартных полей, ckanext-scheming. Региональный сборщик добавляет ckanext-harvest. Эти пакеты в октябре 2026 года тестируются на 2.12, но в документации ckanext-dcat поддержка 2.12 ещё не заявлена. DOI в такую сборку не входит. Список пакетов, кто их ведёт и на каких версиях идут тесты — в главе «CKAN: ядро и расширения».

Запуск и сопровождение. Официальная сборка в Docker: дни или недели до пилота, недели или месяцы, если нужны свои поля и сбор. Каждое расширение проверяют при обновлении ядра. Переход на 2.12 обновляет базу и способ хранения файлов.

Ограничения.

- Без ckanext-dcat портал не попадает в национальный каталог и в поисковики данных: выгрузки DCAT в ядре нет.
- Обязательный DOI на текущем ядре эта платформа не закрывает.
- Закрытый набор скрыт целиком. Схему «описание открыто, файл по заявке» здесь собирают второй системой.
- Регистрация с сайта по умолчанию выключена с версии 2.10.4. Портал, который обновляли со старой версии, часто сохранил прежнюю настройку: роботы заводят учётные записи и публикуют спам.

Примеры. open.canada.ca (Канада), Humanitarian Data Exchange (CKAN 2.11.4), data.gov.au (Австралия, CKAN 2.11.6), data.gov.gr (Греция, CKAN 2.11.3). В октябрьской сверке у австралийского портала были видны ckanext-scheming, ckanext-dcat, ckanext-harvest, ckanext-spatial и xloader. 13 октября 2026 года повторный запрос `status_show` снаружи ответил кодом 403. В тот же день греческий портал вернул 66 имён плагинов: среди них scheming, dcat, harvest, fluent, showcase и проверка наборов высокой ценности, а также свои сборщики ведомств. Как читать такой список — в главе об экосистемах.

Платформу меняют в обе стороны. Американский data.gov работал на CKAN до апреля 2026 года, затем перешёл на собственную разработку; Монголия тоже ушла с CKAN на свой движок. Австралия в 2025 году, наоборот, вернулась на CKAN с Magda, а Греция в 2025–2026 годах перешла на CKAN со своей разработки. Британский data.gov.uk в 2026 году стал National Data Library с новой витриной, прежний каталог оставлен отдельным разделом.

DKAN

Свободная платформа на Drupal, лицензия GPL-2.0. Текущий выпуск модуля — 4.1.5 от 7 октября 2026 года, заявлен для Drupal 10 начиная с 10.2 и для Drupal 11. В реестре 161 портал открытых данных. Страница модуля на Drupal.org 8 октября 2026 года сообщала около 164 сайтов: это отчёт самих установок Drupal, а не подсчёт реестра.

Кому подходит. Организации, у которой сайт уже на Drupal и каталог должен быть его частью.

Кому не подходит. Организации, которая не хочет сопровождать Drupal. Каталогу, которому нужны DOI и выгрузка цитирования.

Возможность	Где
DCAT, API, наборы	в ядре DKAN: файл `data.json`, схема ядра — DCAT-US 1.1
Свои поля	в ядре: схему набора задаёт файл JSON Schema, файл по умолчанию заменяют своим
DOI и отдельный экспорт цитирования	нет
schema.org, сбор чужих каталогов, единый вход	не проверено; вход и права — модули Drupal
Модули сайта, права, языки	экосистема Drupal

Экосистема. Два календаря обновлений: дыру в Drupal и дыру в DKAN закрывают разными выпусками. Состояния набора в документации ветки 2.x — черновик, опубликован, скрыт и в архиве; по умолчанию набор публикуется сразу. Перед установкой 4.1.5 сверяют, что эта документация всё ещё про ту ветку, которую ставите.

Запуск и сопровождение. По сроку это свободное ПО с расширениями: недели или месяцы, не дни. DCAT добавлять отдельным пакетом не нужно, в отличие от CKAN.

Ограничения. Своя схема — замена файла в поставке, а не каталог пакетов вроде ckanext-scheming. DOI и готовой ссылки для менеджера библиографии нет. Сбор чужих каталогов в октябрьской сверке не подтверждён: если портал должен забирать ведомства, сначала смотрят CKAN или Piveau.

Примеры. Каталог данных Medicaid — портал CMS. Файл описаний лежит по адресу `/data.json`.

udata

Свободная платформа data.gouv.fr, лицензия AGPL-3.0, её ведёт команда Open Data Team. Текущий выпуск — 17.13.0 от 13 октября 2026 года. В реестре 15 порталов открытых данных, главный — французский: в сверке 12 октября 2026 года он работал на 17.12.1 и держал 74 376 наборов.

Кому подходит. Порталу, где рядом с наборами живут описания API, обсуждения и сбор чужих каталогов, и есть команда на Python-приложение.

Кому не подходит. Порталу, которому нужна разметка страниц для Google Dataset Search без своей доработки. Организации, которая хочет выбирать функции из каталога расширений, как у CKAN.

Возможность	Где
DCAT	в ядре: по страницам и файлом у набора
Выгрузка всего каталога	в ядре: ночная задача пишет CSV
Каталог API и сбор чужих каталогов	в ядре: на data.gouv.fr 1 265 сервисов данных и 507 источников сбора
Роли	в ядре: администратор, редактор, с версии 15.4.0 (март 2026) ещё и частичный редактор

Возможность	Где
Черновик	в ядре: это закрытый набор
schema.org на странице	нет: в HTML блока разметки нет
Проверка куратором перед публикацией	нет
DOI	нет

Экосистема. Доработки делают в своей установке. Отдельного каталога расширений с чужими сопровождающими, как у CKAN, здесь нет. Подозрительные наборы и обсуждения платформа умеет помечать по списку слов и отдавать модератору.

Запуск и сопровождение. Своя команда или подрядчик, как у другой свободной платформы портала: недели или месяцы до рабочего каталога. data.gouv.fr в день выпуска 17.13.0 ещё отвечал предыдущим номером: крупный сайт обновляют после проверки, а не в день тега.

Ограничения. Сборщик каталогов увидит DCAT. Поисковик, который читает только разметку страницы, набор не увидит, пока разметку не добавят сами. Профиля страны «из коробки», как список спецификаций у Piveau, в этой карточке нет: французский портал — образец устройства, а не готовый национальный профиль для другой страны.

Piveau

Свободная платформа каталогов Fraunhofer FOKUS, построенная вокруг DCAT и сбора чужих каталогов. Компонент каталога piveau-hub-repo выходит под Apache-2.0; текущий выпуск — 4.4.3 от 24 сентября 2026 года. В реестре 19 порталов. На Piveau работает data.europa.eu.

Кому подходит. Национальному или тематическому порталу, который в основном собирает чужие каталоги, если нужный профиль DCAT уже есть в продукте.

Кому не подходит. Региону со своими полями, которых нет в форме ввода: такой профиль дальше сопровождает подрядчик. Организации без команды на несколько сервисов платформы.

Возможность	Где
DCAT в нескольких форматах, SPARQL	в ядре
Сбор чужих каталогов и проверка качества описаний	в ядре, это основная работа платформы
Свои поля по готовым спецификациям	в ядре формы ввода: на октябрь 2026 года документированы DCAT-AP и DCAT-AP.de
schema.org на странице набора	нет: страница на data.europa.eu рисуется скриптом, в HTML разметки нет
DOI	нет

Экосистема. Платформа собрана из нескольких сервисов. Версия 4.4.3 — у компонента каталога, а не единый номер всей поставки. Чужой модуль «как ckanext» сюда не ставят: свой профиль добавляют отдельной работой.

Запуск и сопровождение. Свободная платформа со своей схемой: недели или месяцы. Обновления делает своя команда. На data.europa.eu описание набора в Turtle открывается по адресу API, даже когда страница для человека собрана скриптом.

Ограничения. Профиль, которого нет в списке формы, в короткий список не кладут. Разметки страницы для Google Dataset Search нет. Свои расширения на сервер, как у CKAN, не предусмотрены.

Huwise (OpenDataSoft) и Socrata

Коммерческие облачные порталы открытых данных. В реестре 454 портала, записанных как OpenDataSoft, и 245 порталов Socrata. Свой модуль на сервер поставить нельзя: остаётся то, что есть в тарифе, плюс виджет и API.

30 сентября 2025 года OpenDataSoft переименовалась в Huwise и называет продукт маркетплейсом данных: портал открытых данных и витрина обмена внутри организации и с партнёрами. Платформа и API остались прежними, адреса порталов не менялись. В реестре платформа записана старым именем. Дальше в книге — Huwise, старое имя в скобках. Socrata принадлежит Tyler Technologies и продаётся как Tyler Data & Insights.

Кому подходит. Органу власти без своей команды, которому нужны API, DCAT и таблицы в браузере в ближайшие недели.

Кому не подходит. Организации, которой нельзя выносить данные к поставщику. Каталогу, которому нужен DOI или выгрузка в профиль, которого у поставщика нет.

Возможность	Huwise	Socrata
API, DCAT, хранение таблиц, роли	в ядре сервиса	в ядре сервиса; файл `data.json` по DCAT-US
Свои поля описания	шаблоны полей с переводом	наборы полей с обязательностью и проверкой
Словарь полей таблицы	подпись, описание и тип колонки	то же
schema.org	в ядре, включает администратор	JSON-LD на странице набора
Закрытые наборы	набор виден в каталоге, содержимое — выбранным; запрос доступа включается в настройках	закрытый набор скрыт целиком, запроса доступа нет
Сбор чужих каталогов	сборщики CKAN, Socrata, ArcGIS, ArcGIS Hub, `data.json`, data.gouv.fr и других порталов Huwise; сбор из CSW включают по запросу в поддержку	Catalog Connector забирает `data.json` и публичный ArcGIS Server; набор из `data.json` после синхронизации становится открытым
DOI, свои расширения	нет	нет
Единый вход	SAML и OpenID Connect	OpenID Connect; SAML через поддержку

Запуск и сопровождение. Дни, как у облачного сервиса. Обновления делает поставщик. Сбор чужих каталогов есть в самом сервисе, отдельный модуль ставить некуда.

Ограничения. Уйти позже стоит выгрузки всех наборов и файлов и переноса адресов. Национальный профиль DCAT, которого нет в тарифе, поставщик не допишет как расширение CKAN: его либо включают на пилоте, либо выбирают другую платформу. Смена имени или владельца — повод перечитать договор о выгрузке, а не повод отказываться от сервиса заранее.

Примеры. Витрина поставщика — бывший data.opendatasoft.com, теперь hub.huwise.com.

JKAN

Свободный статический каталог на Jekyll, лицензия MIT. В реестре 48 порталов открытых данных.

Кому подходит. Организации с десятками или несколькими тысячами наборов, если их правят файлами в Git или формой, которая сохраняет правку в тот же репозиторий.

Кому не подходит. Каталогу с ролями, модерацией, заявками на закрытый файл и API поиска.

Возможность	Где
Страницы наборов, файл `data.json`	в ядре шаблона. Файл объявляет схему DCAT-US 1.1 и заполняет короткий список полей
Форма редактора	Decap CMS: правка становится коммитом в Git. На GitHub Pages инструкция просит отдельный сервер входа
API поиска, DOI, роли, своё хранилище с правами	нет. Поиск на странице фильтрует готовый JSON в браузере
schema.org	шаблон страницы разметку не ставит; её добавляют правкой шаблона
Лента изменений	лента репозитория, если сайт собран из Git

Запуск и сопровождение. Часы или дни. Сервер приложений не нужен. Сопровождает тот, кто правит файлы: когда он уходит, сайт умеет собирать следующий человек или каталог замирает.

Ограничения. API поиска и ролей в шаблоне нет. Сбор чужих каталогов делают отдельным скриптом, как на opendata.scot. Что проверить в `data.json` и на странице — в главе о статических сайтах.

Примеры. opendata.scot. Остальные лёгкие варианты — в части «Публикация без портала».

Dataverse

Свободный репозиторий научных данных, лицензия Apache-2.0, проект IQSS Гарвардского университета. Текущий выпуск — 6.12.1 от 8 октября 2026 года. В реестре 257 репозиториев. В карточке реестра указана версия 6.9, Harvard Dataverse в ту же сверку работал на 6.10.1.

Кому подходит. Университету и научному консорциуму, которым нужны DOI, версии, цитирование и модерация, и у которых есть библиотека или вычислительный центр.

Кому не подходит. Порталу открытых данных с ежедневными обновлениями и федерацией DCAT-AP: для этой задачи берут CKAN или Piveau.

Возможность	Где
Файлы, версии, OAI-PMH, строка цитирования, BibTeX и RIS	в ядре
DOI	в ядре: DataCite, с выпуска 6.12.1 ещё и Crossref, а также EZID, Handle и постоянные ссылки установки

Возможность	Где
schema.org	в ядре; с версии 6.10 в страницу встроена краткая разметка Croissant
FAIR Signposting	в ядре с версии 5.14
Свои блоки метаданных	в ядре
Просмотрщики файлов	внешние инструменты: в установку кладут манифест, само приложение может работать на чужом сайте
DCAT	в ядре нет; сторонний экспортёр, механизм экспортёров документация называет экспериментальным
Облачное хранилище	в ядре: S3, Swift, Globus, несколько хранилищ
Единый вход	в ядре: SAML, OpenID Connect, ORCID

Экосистема. Основное в ядре. Снаружи — просмотрщики и дополнительные форматы выгрузки, их ведёт консорциум GDCC. С версии 6.1 просмотрщику рекомендуют короткую подписанную ссылку на файл. Долгоживущий ключ API после версии 6.0 документация не рекомендует: он действует как пароль пользователя. Чужой формат выгрузки с версии 5.14 подключают JAR-файлом внутри установки, и его проверяют при обновлении.

Запуск и сопровождение. Есть официальная инструкция. Первая установка тяжелее статического сайта. С версии 6.8 контейнеры считают готовыми для рабочих установок; несколько серверов приложений в контейнерах не поддерживаются. Договор с DataCite заключают отдельно от установки. Сопровождение — задача библиотеки или вычислительного центра, не разового гранта.

Ограничения.

- Уметь выдавать DOI и выдавать его — разные вещи. В реестре пометка стоит у 172 из 257 установок: у остальных договор с агентством не подключён или пометка снята вручную.
- Страница набора Harvard Dataverse отвечает роботу проверкой AWS WAF, хотя API и OAI-PMH работают. Проверку «открывается ли сайт» делают и браузером, и запросом к API.
- Экспортёр DCAT при обновлении может потребовать доработки. Для портала, который обязаны забирать по DCAT-AP, это слабый путь.

Примеры. Harvard Dataverse. Лимиты бесплатной публикации — в главе о научных репозиториях.

InvenioRDM

Свободный репозиторий, лицензия MIT. Текущий выпуск — v14 от 7 августа 2026 года. Версию v13 поддерживают до 7 февраля 2027 года, v15 в октябре 2026 года была в бете. Политика проекта: одна крупная версия в год, исправления — не меньше года и ещё полгода после следующей. В реестре 74 установки, включая Zenodo.

Кому подходит. Организации, которой нужен репозиторий с DOI, сообществами и закрытыми записями и которая готова обновлять Python-приложение раз в год.

Кому не подходит. Небольшой группе, которой достаточно положить файлы: для этого есть сам Zenodo. Порталу открытых данных.

Возможность	Где
Файлы, DOI, версии, OAI-PMH, DCAT	в ядре
Цитирование	в ядре: стили CSL, BibTeX, CSL-JSON
RIS	нет
Дополнительные поля	настройка ядра, не чужой каталог плагинов
Закрытые записи и запрос доступа	в ядре
Облачное хранилище	в ядре с версии v13: S3
FAIR Signposting	в ядре с версии v12
Единый вход	в ядре: ORCID, OpenID Connect, SAML
Croissant	нет
schema.org	в ядре

Экосистема. Документация выстраивает доработки от дешёвой к дорогой: настройка, файлы в ожидаемых местах, готовый модуль, свой пакет в папке установки, правка ядра с отдачей правки в проект, своя копия чужого репозитория. Логотип меняют настройкой. Свою копию ради логотипа заводят только если готовы переносить каждое обновление руками. Подробнее — в главе об экосистемах.

Запуск и сопровождение. Официальная инструкция есть, первая установка тяжелее статического сайта. Рядом стоит выбор опубликовать в Zenodo и свою установку не заводить. Для небольшой группы он обычно лучше. Крупная установка отстаёт от выпуска так же, как Harvard Dataverse: это нормально внутри окна поддержки и становится риском, когда окно закрылось.

Ограничения. Документация метаданных опирается на схему DataCite 4.3, а текущая схема — 4.7 от марта 2026 года. Новые поля, например связь с проектом, перед выбором проверяют по этой документации. RIS из ядра не выгружается. Свою установку имеет смысл заводить, когда библиотека модерирует много групп и нужен свой префикс DOI.

Примеры. Zenodo, CaltechDATA. Лимиты Zenodo — в главе о научных репозиториях.

DSpace

Свободный репозиторий, лицензия BSD. В реестре 829 научных репозиториев: это в первую очередь репозиторий публикаций, куда кладут и данные. Текущий выпуск — 10.1 от 14 сентября 2026 года; в тот же день вышли исправления 9.4 и 8.5. В DSpace 10 появился тип записи «набор данных». В карточке реестра указана версия 9.2, DSpace@MIT работал на 8.2.

Кому подходит. Библиотеке, у которой DSpace уже хранит публикации и туда же кладут данные.

Кому не подходит. Новому репозиторию данных, если публикаций нет: Dataverse и InvenioRDM ближе к данным. До версии 10.1 включительно DSpace не закрывает выгрузку цитирования, поэтому в короткий список репозитория данных его не ставят.

Возможность	Где
Файлы, Dublin Core, OAI-PMH	в ядре
DOI	в ядре, включается настройкой; по умолчанию идентификатор записи — Handle, адрес строится от него
Свои поля и формы ввода	настройка ядра: реестр схем метаданных и файл `submission-forms.xml`
Готовая ссылка и выгрузка в BibTeX и RIS	нет до 10.1 включительно: в ядре только метатеги для Google Scholar и Zotero; кнопку «Cite» переносят из DSpace-CRIS, перенос не завершён
Импорт записей из BibTeX, RIS и EndNote	в ядре
FAIR Signposting	в ядре с версии 7.6
schema.org и DCAT	в ядре не найдены
Облачное хранилище	в ядре: S3, несколько хранилищ
Тема оформления и часть интеграций	свой код в модуле `additions` поверх классов ядра

Экосистема. Чужого каталога плагинов нет. Свою правку заново накладывают на новую версию. Это дешевле, чем писать репозиторий с нуля, и дороже, чем включить функцию в Dataverse.

Запуск и сопровождение. Официальная инструкция, первая установка тяжелее статического сайта. Библиотеке, у которой DSpace уже стоит, обновление дешевле новой установки ради данных. Новому репозиторию данных без уже стоящего DSpace этот срок обычно не окупается.

Ограничения. В реестре пометка DOI стоит у 218 из 829 установок: идентификатор включают отдельно, и многие репозитории живут на одном Handle. Метатеги для Zotero — это задел, а не кнопка «скопировать ссылку». Тип записи «набор данных» появился в версии 10; установка на 8.x его не имеет, пока её не обновят.

Примеры. DSpace@MIT.

Figshare

Облачный сервис Digital Science: общий сайт figshare.com и репозитории организаций. В реестре 288 научных репозиториев. Свой модуль на сервер поставить нельзя.

Кому подходит. Университету или автору, которым нужен DOI без своей установки и которые принимают лимиты и тариф.

Кому не подходит. Порталу открытых данных. Организации, которой нельзя отдавать файлы поставщику. Массиву больше лимита бесплатного тарифа, если разовая плата за объём не заложена.

Возможность	Где
Файлы, DOI, строка цитирования, BibTeX и RIS	в ядре сервиса
API	в ядре сервиса
Свои расширения	нет

Запуск и сопровождение. Дни, как у облачного сервиса. Обновления делает поставщик. На 12 октября 2026 года бесплатно было 20 ГБ личного места и файл до 20 ГБ. Figshare+ берёт разовую плату 1 225 долларов за каждые 250 ГБ, файлы до 5 ТБ. В реестре пометка DOI стоит у 266 из 288 установок.

Ограничения. Уйти можно тем, что поставщик отдаёт выгрузкой. Автоматический сбор коммерческих репозиториев ограничен: в главе о научных репозиториях это названо отдельной трудностью. Тариф и место хранения проверяют в договоре, а не по чужому пересказу.

GeoNetwork

Свободный каталог описаний геоданных, лицензия GPL-2.0. Текущий выпуск линейки 4.4 — 4.4.12 от 8 июля 2026 года, в нём закрыты уязвимости предыдущих версий 4.4. В реестре 762 геопортала. В карточке реестра указана версия 4.4.9.

Кому подходит. Инфраструктуре пространственных данных, где нужны ISO 19115, CSW и сбор карточек из других каталогов.

Кому не подходит. Порталу таблиц и реестров: там проще CKAN или облачный сервис. Организации, которой нужен один сайт «карта и каталог» без отдельного сервера слоёв: ближе GeoNode.

Возможность	Где
ISO 19115, ISO 19139, ISO 19115-3, Dublin Core, каталог объектов ISO 19110	в ядре
CSW, OAI-PMH, сбор карточек (CSW, OAI-PMH, сервисы OGC, коллекции STAC)	в ядре
DCAT, в том числе GeoDCAT-AP	в ядре
schema.org и карта сайта	в ядре: JSON-LD на странице записи (поля Croissant для ISO 19115-3 — в ветке будущего выпуска 4.4.13), карта сайта указана в `robots.txt`
DOI	в ядре, нужно включить и подключить DataCite
Готовая ссылка на запись	в ядре: HTML, текст, RIS, BibTeX; состав строки задан в файле XSL схемы
Ограничение скачивания при открытом описании	в ядре: отдельное право на файлы, загруженные в GeoNetwork
Облачное хранилище	в поставке: модули S3, JCloud и CMIS
Версии записи	нет; при согласовании есть рабочая копия
OGC API Records	отдельный микросервис, не страница каталога
Профили метаданных под страну	плагины схем

Экосистема. Схемы ISO из поставки — это ядро. Профиль страны — отдельный плагин: каталог схем и преобразований, иногда свой JAR. Плохой плагин ломает весь каталог, об этом предупреждает руководство администратора. Ставят ветку, названную вашей линейкой GeoNetwork, и проверяют, что такая ветка в репозитории есть. Две копии британского профиля GEMINI разобраны в главе об экосистемах.

Запуск и сопровождение. Официальная инструкция, в том числе Docker. Первая установка тяжелее статического сайта. Рядом почти всегда нужен сервер карт: GeoNetwork описывает слой, GeoServer или другой сервер его отдаёт. Профили страны проверяют при каждом обновлении.

Ограничения.

- Пометка DOI в реестре значит «платформа умеет». В DataCite 12 октября 2026 года DOI вели примерно на 45 установок GeoNetwork, а пометка в реестре стояла у 555. Перед требованием «у всех наших слоёв будет DOI» смотрят договор с DataCite, а не клетку реестра.
- Номер на живом сайте и номер, который стоит скачивать, сверяют отдельно. Каталог Ifremer Sextant 13 октября 2026 года показывал 4.4.13 — новее объявленного выпуска 4.4.12.
- Портал открытых данных с федерацией DCAT-AP как главной задачей здесь не собирают: DCAT геокаталог отдаёт, ежедневную витрину таблиц — нет.

Примеры. Nationaal Georegister — национальный каталог Нидерландов. Каталог Европейского агентства по окружающей среде и национальные каталоги ЕС разобраны в главе о геопорталах.

GeoNode

Свободная геоплатформа OSGeo, лицензия GPL версии 2 или новее. В реестре 770 геопорталов. В главах книги сверен выпуск 5.1.0 (2026 год): карта и каталог собраны вместе.

Кому подходит. Организации, которая публикует файлы вручную и хочет один сайт: загрузил файл — получил слой, карточку, карту и ссылку на скачивание.

Кому не подходит. Национальной инфраструктуре, где каталог и сервер карт живут отдельно, а описания уходят в CSW с профилем страны: там GeoNetwork. Маленькому городу свой сервер обычно вреден: его некому обновлять после проекта, на деньги которого его поставили.

Возможность	Где
Загрузка файла, карта, скачивание, API	в ядре; слой отдаёт GeoServer, карточки по CSW — pycsw
ISO 19115 и Dublin Core по CSW	в ядре, через pycsw
Документы, истории на карте, дашборды	в ядре версии 5.1
DCAT и файл `data.json`	нет: в коде 5.1 их не нашли
schema.org на странице	нет в сверке 13 октября 2026 года
DOI	нет

Экосистема. Свои данные и витрина — одна установка. Внутри уже стоят GeoServer и pycsw, их обновляют вместе с GeoNode. Отдельный каталог плагинов схем, как у GeoNetwork, для обычной публикации не нужен. Национальный профиль ISO сюда не «включают галочкой».

Запуск и сопровождение. Свободная платформа со своим сервером: недели или месяцы, дальше — своя команда. В реестре не действуют 11% геопорталов на GeoNode. Частая история — геопортал построили по донорской программе, программа кончилась, сервер выключили.

Ограничения. Описания в DCAT эта платформа не отдаёт. Поисковик открытых данных и национальный портал на DCAT-AP её каталог сами не заберут: нужна вторая выгрузка или второй каталог. Для слоёв, которые уже лежат в региональной инфраструктуре, свой GeoNode дублирует сервер, который кто-то уже сопровождает.

Примеры. Национальный геопортал Казахстана map.gov.kz: 12 октября 2026 года его API без входа отдавал описания 378 наборов.

ArcGIS Hub

Облачный каталог Esri, коммерческая подписка. В реестре 3 464 геопортала. Сайты одновременно похожи на геопортал и на портал открытых данных: слой можно скачать файлом, каталог отдаёт DCAT. Порталом открытых данных реестр записывает Hub редко — 19 раз. Свой модуль на сервер поставить нельзя.

Кому подходит. Организации, у которой карты уже в ArcGIS и нужно быстро дать им страницу со скачиванием.

Кому не подходит. Организации, которая не хочет зависеть от лицензий Esri. Каталогу, которому нужен DOI.

Возможность	Где
Каталог, файлы и слои, OGC API Records	в ядре сервиса
DCAT	в ядре: по умолчанию ленты DCAT-US 1.1 и DCAT-AP 2.1.1; ленты DCAT-US 3.0 и DCAT-AP 3.0.0 включаются отдельно
Стандарт описания	один стиль на организацию: ISO 19139, ISO 19115-3, INSPIRE, FGDC, североамериканский профиль или Dublin Core+ (по умолчанию у организаций, созданных после июля 2025 года)
Описания полей слоя	в ядре: описание, тип значения, списки допустимых значений
DOI, свои поля сверх выбранного стандарта, свои расширения	нет
schema.org	в сверке октября 2026 года на странице не найден
Сбор чужих каталогов	не проверено

Запуск и сопровождение. Дни, как у облачного сервиса. Обновления делает поставщик. Уйти позже стоит выгрузки и переноса адресов.

Ограничения. Один стиль метаданных на всю организацию: второму ведомству с другим стандартом нужен другой сайт или другая платформа. Лента DCAT-AP 3 сама не включается. Набор, который видно только на карте и нельзя скачать, каталогом открытых данных не считается — это проверяют на пилоте.

Примеры. Каталоги организаций собраны на hub.arcgis.com. Портал округа Колумбия opendata.dc.gov — пример сайта, который отдаёт и карту, и DCAT.

PxWeb

Свободная платформа статистических таблиц Статистики Швеции, лицензия Apache-2.0. В реестре 110 каталогов показателей, среди них Statbank Denmark. Последний выпуск старой линии — 2025.v1 (февраль 2025 года). Новое поколение — API PxWebApi 2 (у Статистики Швеции с октября 2025 года, выпуск 2.7.0 — сентябрь 2026) и интерфейс PxWeb 2 выпуска 2.12.0 от 7 октября 2026 года. Интерфейс берёт данные из этого API.

Кому подходит. Статистической службе, которая публикует показатели с разрезами и пока отдаёт их людям и программам в JSON-stat, CSV и PX.

Кому не подходит. Каталогу файлов. Службе, которой нужен обмен по SDMX со структурами и справочниками: его закрывают .Stat Suite или SDMX-RI, часто рядом с PxWeb, а не вместо него.

Возможность	Где
API	в ядре; PxWebApi 2 принимает запросы в адресе и отдаёт JSON-stat 2, CSV, PX, Excel, HTML; Parquet — в бета-версии
Список всех таблиц и выборка изменённых за N дней	в ядре PxWebApi 2
SDMX	только файл SDMX-ML 2.0 в старом API; в PxWebApi 2 SDMX среди форматов нет
DCAT, schema.org, DOI	нет
Реестр структур SDMX	нет

Запуск и сопровождение. Свободное ПО без своей схемы полей: дни или недели до пилота. Новые установки имеет смысл сразу сажать на PxWebApi 2. Переход меняет адреса запросов; старые ссылки пользователей переводит конвертер Статистики Швеции. Подробнее — в главе о каталогах показателей.

Ограничения. Выгрузку DCAT реестр нашёл меньше чем у десятка каталогов показателей на любых платформах. Чтобы таблицы попали на портал открытых данных, описания делают отдельно. Полноценной службы SDMX в PxWeb нет, и новое поколение API этот пробел не закрыло.

.Stat Suite

Свободная платформа сообщества SIS-CC, которое ведёт ОЭСР. В реестре 40 каталогов показателей. Она собрана вокруг SDMX: обозреватель данных (Data Explorer, выпуск 27.7.0 — июль 2026 года), инструмент загрузки и служба SDMX. Служба внутри — SDMX-RI Евростата под лицензией EUPL. Поставившая .Stat Suite служба на деле ставит и её.

Кому подходит. Статистической службе, которой нужен обмен с Евростатом, МВФ или ОЭСР и есть консультанты или своя команда, знающая SDMX.

Кому не подходит. Службе, которой пока хватает таблиц для людей и JSON-stat: там быстрее PxWeb. Организации, которая хочет «портал показателей за недели» без структур и справочников.

Возможность	Где
SDMX API, реестр структур	в ядре; структуры можно вести и в Fusion Metadata Registry
Обозреватель данных отдельно от службы	в ядре поставки: это разные модули
DCAT	не проверено
DOI	нет

Запуск и сопровождение. Внедрение тяжёлое, без знания SDMX его ставят редко. Обновляют несколько модулей, а не один контейнер. Образец интерфейса — OECD Data Explorer. На той же платформе работают ФАО, МОТ и статистические службы Австралии, Новой Зеландии, Италии и Швейцарии.

Ограничения. Это не замена PxWeb для простой витрины и не каталог файлов. Сочетание «PxWeb для людей, .Stat Suite для обмена» — обычная схема статистической службы. Склеивать её с каталогом микроданных на NADA в одну программу не стоит: стандарты и уровни доступа разные. Разбор — в примере статистической службы.

NADA

Свободный каталог микроданных, лицензия MIT, Всемирный банк и сеть IHSN. Текущий выпуск — 5.7.1 от 4 октября 2026 года (5.6 — июль 2026, 5.7 — сентябрь). В карточке реестра указана версия 5.5. В реестре 158 каталогов микроданных, 47 из них не открываются.

Кому подходит. Статистической службе и архиву обследований, которым нужны уровни доступа, заявки и описание в DDI.

Кому не подходит. Порталу открытых таблиц. Университетскому архиву, которому нужны DOI и сдача описаний в европейский каталог CESSDA по OAI-PMH: там ближе Dataverse.

Возможность	Где
Каталог исследований, поиск по переменным, семь типов доступа, заявки, API	в ядре
Выгрузка DDI Codebook, разметка schema.org (JSON-LD)	в ядре
DOI	поле для DOI, выпущенного отдельно в DataCite; сама NADA DOI не регистрирует
Экспорт цитирования	в ядре, через DataCite, если у исследования есть DOI
OAI-PMH, DCAT	нет
Показатели, документы, таблицы и геоданные	в ядре с выпуска 5.6; службы SDMX и сервисов слоёв нет
Шаблоны описаний	с выпуска 5.7 — шаблоны редактора метаданных Всемирного банка

Экосистема. Отдельного каталога расширений мало. Для администраторов есть клиенты API для R и Python (NADAR и PyNADA). Описания пишут в редакторе метаданных, а NADA их показывает. Замену закрытому Nesstar как раз ищут здесь или в Dataverse; сам Nesstar для нового каталога не выбирают.

Запуск и сопровождение. Дни или недели до пилота, если не писать уровни доступа заново. Главный риск — не установка, а сопровождение после проекта: 47 из 158 каталогов в реестре не открываются. Это старые установки статистических служб, часто без HTTPS, а не след снятого Nesstar. Подробнее — в главе о каталогах микроданных.

Ограничения.

- В европейскую федерацию архивов описания попадают только через промежуточную выгрузку: CESSDA забирает DDI по OAI-PMH, а этого протокола в NADA нет.
- DOI регистрируют сами в DataCite и вписывают в поле. Пока поле пустое, экспорт цитирования молчит.
- Выпуск 5.6 умеет описать показатель, но не заменяет каталог показателей. В библиотеке Всемирного банка все записи остаются обследованиями. Показатели и микроданные держат разными каталогами.

Примеры. Каталог микроданных Всемирного банка, сводный каталог IHSN.

Платформы без отдельной карточки

udata, Piveau, GeoNode, .Stat Suite и Figshare разобраны выше. Ниже — платформы отдельных ниш. Лицензию и срок перед решением смотрят в документации так же, как у карточек. Число установок — по той же выгрузке реестра, если не указан другой источник.

Порталы и агрегаторы

- **Magda** — свободный (Apache 2.0) австралийский каталог-агрегатор: собирает описания из многих источников в один поиск. В 2019–2025 годах давал поиск data.gov.au. Сейчас на нём, например, казахстанский Ashyq Data. Тяжёл в сопровождении и, по оценке 2024 года, окупается только на десятках тысяч наборов.
- **EntryScape** — коммерческая платформа шведской компании MetaSolutions: каталог по DCAT-AP с публикацией на своём сайте и в национальный dataportal.se. В реестре 99 порталов открытых данных, почти все в Швеции.

Геоплатформы (подробнее о ролях — в главе о геопорталах)

- **pycsw** — свободный (MIT) лёгкий каталог описаний геоданных: CSW, OGC API Records, STAC API, OAI-PMH. Встроен в GeoNode, подключается к CKAN; отдельных установок 44.
- **Oskari** — свободная (MIT или EUPL) платформа картографических витрин, её координирует Национальная земельная служба Финляндии. Показывает слои из GeoServer, GeoNetwork и других сервисов. 23 установки, 21 из них в Финляндии, включая национальный Paikkatietoikkuna.
- **TerriaJS** — свободная (Apache 2.0) библиотека карт-витрин, созданная в CSIRO. Собирает на одной карте слои WMS, WFS, ArcGIS и наборы из CKAN, Socrata, OpenDataSoft. 75 установок, например Digital Earth Africa Map. Национальная карта Австралии на TerriaJS закрыта 27 июня 2025 года.
- **QGIS Web Client 2 (QWC2)** — витрина карты поверх чужих сервисов. 159 установок, больше всего в Финляндии и Германии. Своих данных почти не хранит.

- **GeoBlacklight** — свободный (Apache 2.0) поиск по картам и геоданным на основе Blacklight. Его ведут университетские библиотеки, пример — геопортал альянса Big Ten. 23 установки.

Наука, статистика и микроданные

- **ERDDAP** — свободный сервер научных данных NOAA, класс «сервер протокола»: выгрузка частей сеточных и табличных наборов в разных форматах, графики и карты. Широко используется в океанологии и метеорологии; 175 установок.
- **GBIF IPT** — свободный (Apache 2.0) инструмент публикации данных о биоразнообразии в сеть GBIF, отраслевая система. В GBIF на 12 октября 2026 года зарегистрировано 392 установки; в реестре 11 российских, их держат институты и университеты.
- **WEKO3** — свободный (BSD-3) репозиторий Национального института информатики Японии на основе Invenio 3. На нём облачный сервис JAIRO Cloud: на конец 2025 финансового года им пользуются 822 организации. В реестре 161 японский репозиторий на WEKO3. О научных репозиториях — в отдельной главе.
- **InstDB** — репозиторий научных данных Центра компьютерных сетей и информации Академии наук Китая из набора FairStack: проверка перед публикацией, открытый и ограниченный доступ, идентификаторы DOI и CSTR. 83 установки, почти все в Китае, например репозиторий Центра научных данных АН Китая.
- **OpenSDG** — свободный (MIT) статический сайт показателей целей устойчивого развития. 72 установки, среди пользователей — Армения, Казахстан, Киргизия. Подробнее — в главе о статических сайтах.
- **PxStat** — свободная (MIT) платформа показателей статистической службы Ирландии: таблицы в JSON-stat, PX, CSV и XLSX, API, график выпусков. Почти не тиражирована: в реестре 2 установки, главная — data.cso.ie; последний выпуск — февраль 2025 года.
- **Redatam** — ПО отдела народонаселения ЭКЛАК ООН (CELADE) для обработки и публикации микроданных переписей и обследований, развивается с 1985 года. Распространяется бесплатно; текущее поколение RedatamX (с 2024 года) работает в Windows, Linux и macOS и как пакет для R. Формат базы собственный, без опубликованного описания. Класс — каталог микроданных; основная область — Латинская Америка; 38 каталогов в реестре. См. главу о каталогах микроданных.
- **OBiBa Mica** — свободный (GPL-3.0) портал эпидемиологических исследований и консорциумов: описания исследований, словари переменных, заявки на доступ к данным. Ближе к каталогу микроданных; 27 установок.

Когда открыть карточку заново

Карточку решения не переписывают от каждого тега. Её открывают, когда меняется ответ на вопрос «функция всё ещё в проверенной сборке».

- Вышла следующая минорная версия ядра, и тесты нужного расширения эту версию не включают. Для CKAN ближайший такой рубеж — уход 2.11 из пары, которая ещё получает исправления.
- Закрылось окно поддержки. У InvenioRDM для v13 это 7 февраля 2027 года.

- Облачный поставщик сменил имя, владельца или состав выгрузки. Так уже было с OpenDataSoft и с Socrata.
- Живой сайт отстаёт от выпуска дальше окна поддержки, а человека на обновление нет.
- Обязательная функция переехала: появился национальный профиль, которого нет в форме Piveau, или ведомство потребовало DOI.

Шаблон записи — в главе «Как обосновать решение».

Источники

- Документация CKAN 2.12, политика поддержки версий, объявление о выпуске 2.12 — 2.12 назван последним выпуском линейки 2.x
- Модуль DKAN на Drupal.org — выпуск 4.1.5 от 7 октября 2026 года; лицензия GPL-2.0
- Журнал изменений udata — 17.13.0 от 13 октября 2026 года; лицензия AGPL-3.0
- Лицензия piveau-hub-repo, Apache-2.0, журнал изменений — выпуск 4.4.3 от 24 сентября 2026 года
- Лицензия GeoNetwork, GPL-2.0, выпуск 4.4.12 от 8 июля 2026 года
- Лицензия GeoNode — GPL версии 2 или новее
- Политика лицензий InvenioRDM, MIT, политика поддержки
- Объявление о переименовании OpenDataSoft в Huwise, 30 сентября 2025 года; настройки безопасности портала Huwise; сборщики Huwise; Catalog Connector у Socrata
- Объявление о новом каталоге data.gov — уход с CKAN в апреле 2026 года
- Link Digital, перевод data.gov.au на CKAN, 2025; GDS, What's changing on data.gov.uk and why, 2026
- Версии и расширения data.gov.au и data.gov.gr — по ответу `/api/3/action/status_show`, октябрь 2026 года. Повторная сверка 13 октября 2026 года: data.gov.gr — CKAN 2.11.3 и 66 имён плагинов, data.humdata.org — CKAN 2.11.4 и 34 имени, data.gov.au ответил кодом 403. Уход opendata.gov.mn с CKAN — И. Бегтин, Telegram, 25.08.2026
- Число установок — выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. Запросы — в приложении.
- Сравнительная таблица

ЧАСТЬ VI.

Как выбрать каталог

От задач к набору функций, от функций к платформе или собственной разработке — с обоснованием каждого решения.

1. Маршрутная карта выбора
2. Принципы выбора набора функций
3. От функций к платформе
4. Как обосновать решение
5. Свои силы, свободное ПО или облачный сервис
6. Типовые сценарии
7. Чек-лист требований
8. Разобранные примеры

Маршрутная карта выбора

Глава отвечает на четвёртый вопрос книги: какой каталог данных выбрать под задачи его создателя. Выбор идёт в три этапа. Сначала по содержимому определяется **тип каталога**. Затем обязательства и режим доступа задают **стандарты и интерфейсы**. Наконец ресурсы, существующая инфраструктура и масштаб определяют **модель создания и платформы**. Если уже понятно, кто владелец — министерство, город, университет, служба, — стартовую строку даёт глава «Типовые сценарии». Маршрут ниже нужен, когда тип ещё не ясен или строка сценария не подошла.

Пройти маршрут

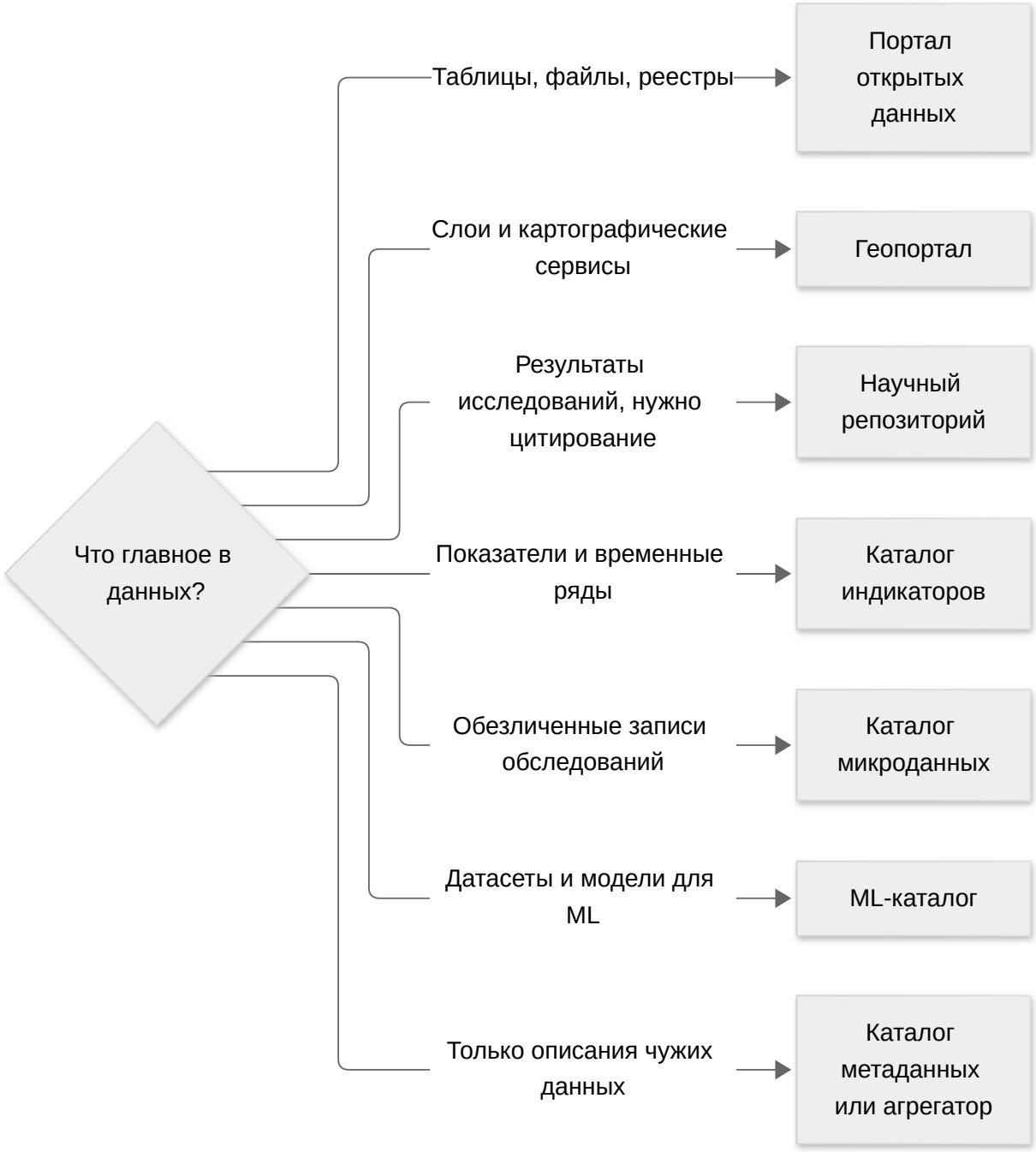
Ответьте на восемь вопросов. Рекомендация — отправная точка для сравнения, а не окончательный выбор. У каждой платформы, каждого стандарта и каждого интерфейса в ответе есть короткое обоснование. Отмеченные возможности сопровождаются пометкой, лежат ли они в ядре платформы или в расширении. Подробный метод — от задач к функциям, от функций к платформе, с обоснованием каждого решения — описан в главах «Принципы выбора набора функций», «От функций к платформе» и «Как обосновать решение».

> Интерактивная маршрутная карта работает в онлайн-версии книги: https://datacatalogs.begtin.tech/choice/route-map/. Ниже — схемы маршрута и логика рекомендаций.

Схема маршрута



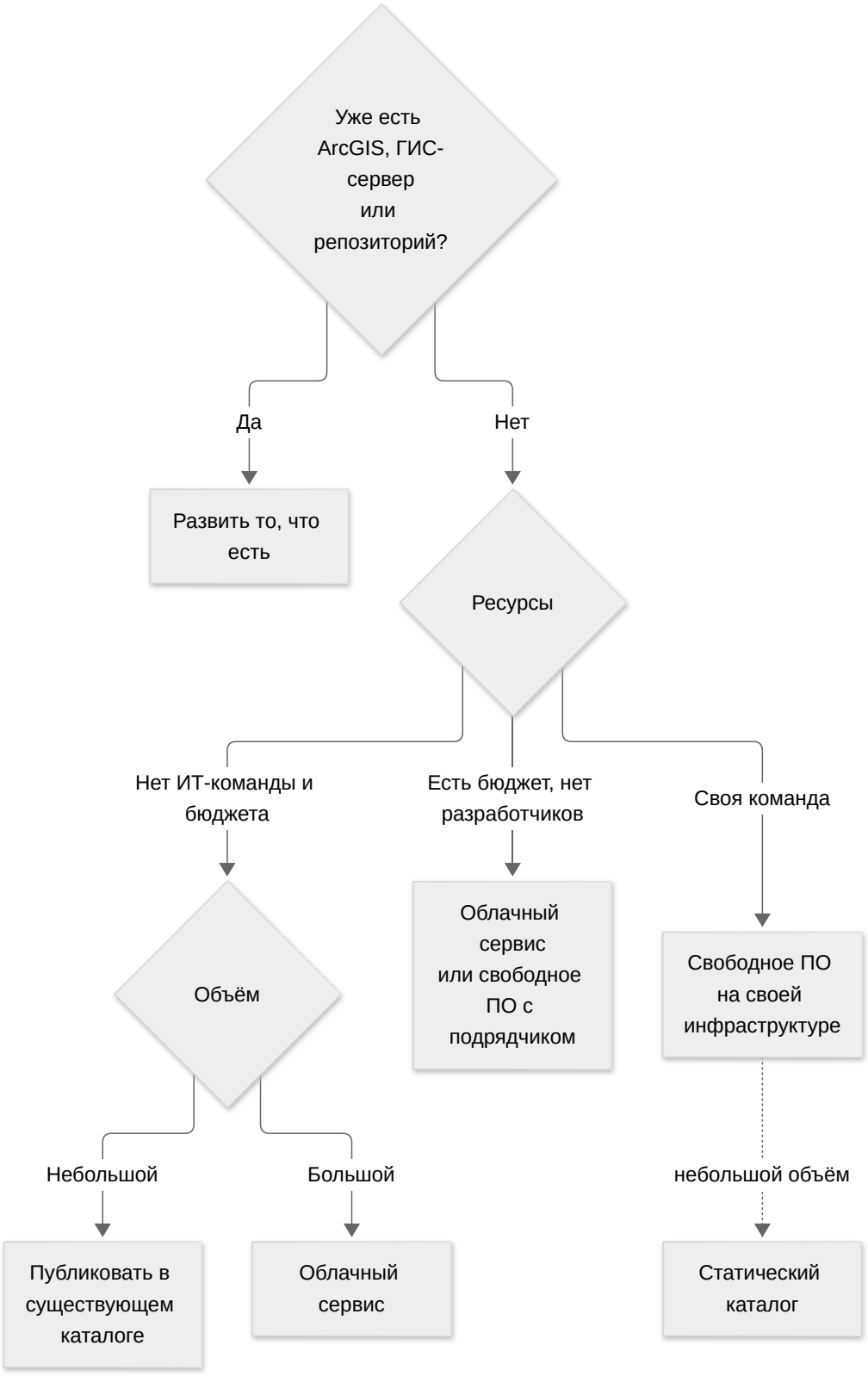
Этап 1. Тип каталога по содержимому



Этап 2. Стандарты и интерфейсы

Отдельной схемы у этого этапа нет: стандарты задаёт тип каталога и обязательства владельца — INSPIRE, сбор вышестоящим порталом, DOI, SDMX, персональные данные. Какие стандарты и интерфейсы обязательны для какого типа, сведено в карте возможностей, а для типовых владельцев — в таблице сценариев.

Этап 3. Модель создания



Как устроена логика

- Рекомендация совпадает со схемами выше.

- Один из вариантов ответа — «не создавать свой каталог». Для малых владельцев без ИТ-команды это часто лучший выбор.
- Варианты, которые стоит сравнить с рекомендацией, показываются рядом с ней.

Источники

- Принципы выбора набора функций
- От функций к платформе
- Как обосновать решение
- Чек-лист требований
- Типовые сценарии

Принципы выбора набора функций

Список функций каталога собирают из задач, а не из рекламы платформы и не из чужого технического задания. У каждой функции в итоговом списке пять пометок: чья задача за ней стоит, насколько она обязательна, почему, как её проверить на живом сайте и когда к решению вернуться. Такой список можно положить в техническое задание и по нему сравнивать платформы.

Маршрутная карта даёт первый ориентир за несколько минут, карта возможностей — что вообще бывает и что типично для вашего типа каталога. Эта глава — способ применить обе карты к своему случаю.

Коротко:

- **Начните с того, что уже есть.** Часть задач закрывают существующие сайты, сервисы и чужие каталоги. Новый каталог часто нужен, чтобы упорядочить уже опубликованное, а не чтобы публиковать с нуля.
- **У каждой функции есть источник**: задача владельца, задача пользователя или внешнее обязательство. Функция без источника — кандидат на вычёркивание, даже если она есть «у всех».
- **Минимум обязателен всем**, остальное включают тип каталога и конкретные обязательства.
- **«Обязательно» проверяют вопросом**: какая задача не выполнится в день запуска без этой функции и кто это заметит. Нет ответа — значит «желательно».
- **У функции есть цена**, и она зависит от того, где функция живёт: в ядре платформы, в расширении, в своей доработке или за пределами портала.
- **Отказ записывают так же, как выбор**, вместе с событием, после которого к нему вернутся.

Возможность, функция, требование

В книге три близких слова, и их стоит различать.

- **Возможность** — то, что каталоги вообще умеют: отдать описания в DCAT, выдать DOI, закрыть файл при открытом описании. Возможностям посвящена часть «Возможности каталогов».
- **Функция** — возможность, которую вы выбрали для своего каталога, потому что за ней стоит чья-то задача.
- **Требование** — функция, записанная в **техническое задание** (документ, по которому подрядчик делает или настраивает систему), вместе с уровнем обязательности, обоснованием и способом проверки.

Глава превращает задачи в функции, а функции — в требования. Международный стандарт инженерии требований ISO/IEC/IEEE 29148:2018 перечисляет свойства хорошего требования. Для каталога важнее всего четыре:

- **необходимое** — за ним стоит чья-то потребность, и её можно назвать;
- **говорит «что», а не «как»** — «национальный портал забирает наши описания в профиле DCAT-AP» — требование, «поставить ckanext-dcat» — уже решение, его принимают на следующем шаге;
- **единичное** — одна функция в одном пункте: «удобный современный поиск с API и картой» — это три пункта, а не один;
- **проверяемое** — можно показать, выполнено оно или нет.

Шаг 1. Опишите, что уже есть

До списка функций составьте опись: какие данные организация уже публикует и где. Раздел «Документы» на сайте, таблицы во вложениях к новостям, API информационных систем, картографический сервер, страницы в чужих каталогах, выгрузки, которые отправляют по запросу. По каждому пункту запишите, кто этим пользуется и что сломается, если это перенести.

Опись даёт две вещи. Во-первых, часть задач уже решена: слои живут в картографическом сервере, научные данные лежат в Zenodo, статистика отдаётся через API. Каталогу остаётся их описать и связать, а не повторить. Во-вторых, видно, что нельзя ломать: адреса, на которые ссылаются, и программы, которые забирают файлы по расписанию.

Пример — Казахстан. В феврале 2023 года национальный портал открытых данных data.egov.kz выдавал данные только после входа по индивидуальному идентификационному номеру через государственную службу идентификации, то есть открытым по сути не был. При этом общедоступных машиночитаемых данных у органов власти было больше, чем на портале: открытый API статистической системы «Талдау» без авторизации, программный интерфейс сайта правительства gov.kz, картографические сервисы геологической службы на ArcGIS. В такой ситуации главная работа портала — систематизировать уже опубликованное, и функции у него другие: описывать и собирать чужие сервисы, а не принимать файлы вручную. В апреле 2023 года обязательную регистрацию на портале отменили, но стали отдавать не больше 100 записей за раз — об этом в главе об антипаттернах. Документация API «Талдау» в октябре 2026 года открыта.

Если опись показывает, что задачи уже закрыты, свой каталог может оказаться не нужен. Когда так бывает — в главе «Когда портал не нужен».

Шаг 2. Соберите задачи

У функции три возможных источника:

- **задача владельца**: отчитаться по закону, выполнить условия гранта, продать сервис, перестать отвечать на одни и те же письма;
- **задача того, кто данными пользуется**: скачать таблицу, сослаться в статье, забрать описания программой, построить приложение;
- **внешнее обязательство**: национальный портал собирает описания в DCAT, журнал требует DOI, инфраструктура пространственных данных требует каталог слоёв.

Где брать задачи:

- **Письма и обращения** с просьбой дать данные. Каждая просьба — задача пользователя, повторяющаяся — кандидат в обязательные.
- **Статистика существующего сайта**: что скачивают и что ищут, но не находят. Запросы в поиске, которые ничего не вернули, — готовый список пробелов.
- **Разговоры с настоящими пользователями** каждой группы: аналитиками, исследователями, разработчиками. Разговоры можно заканчивать, когда ответы начинают повторяться. Чего ждут разные группы — в главе «Зачем создают каталоги и кто ими пользуется».
- **Перечень обязательств**: закон и подзаконные акты, требования вышестоящего портала к профилю описаний, условия грантов, отраслевые инфраструктуры вроде INSPIRE или обмена SDMX. Каждое обязательство записывают со ссылкой на документ.
- **Задачи редакции**: кто будет заводить описания, сколько их, как часто обновлять, откуда брать. Эту группу забывают чаще всего, а от неё зависят роли, массовая загрузка описаний и сбор из учётных систем.

Отдельно запишите задачи программ: вышестоящего портала, поисковиков данных, сборщиков, ИИ-агентов. Писем они не пишут, поэтому их задачи формулируют за них. Обычно это и есть минимум из следующего шага.

Задачу записывают по форме «кто — что делает — зачем». «Аналитик министерства каждый месяц сравнивает качество воздуха по районам» — задача. «Нужен современный портал» — нет: непонятно, кто и что на нём будет делать.

Шаг 3. Переведите задачу в функцию

Одна задача обычно даёт несколько функций, и не всегда тех, что приходят в голову первыми.

Задача	Какие функции из неё следуют	Что часто просят вместо
Журналист раз в год скачивает исполнение бюджета и сравнивает с прошлым годом	файл CSV или XLSX, лицензия, дата обновления, прошлые годы по постоянным адресам	предпросмотр с графиками
Национальный портал собирает описания региона	выгрузка в профиле DCAT-AP по постоянному адресу, лицензия у каждого набора	таблица наборов, которую раз в квартал отправляют почтой
Разработчик приложения показывает ежечасные замеры воздуха	файл или API с постоянным адресом, обновление по расписанию, описание полей	«современный API» без описания, что он отдаёт
Аспирант ссылается на данные в диссертации	постоянный адрес набора и файла; DOI и версии — только если данные меняются, а сослаться нужно на конкретное состояние	DOI на каждый набор
Грант обязывает положить данные к статье	DOI, версия, строка цитирования — часто в общем научном репозитории	свой репозиторий в каждой лаборатории
Поисковик данных находит набор	разметка schema.org, карта сайта, описания без регистрации	продвижение портала в соцсетях
Двадцать отделов публикуют сами	организации, роли, проверка перед публикацией	одна учётная запись на всех
Ведомство отчитывается о числе наборов	это не задача пользователя, функции из неё не выводят	дробление одного ряда на сотни наборов ради счётчика

Если вместо задачи приходит готовое решение — «поставить CKAN», «сделать как у соседей», «нужен API», — спросите, какую задачу оно решает. За словом «API» могут стоять три разные функции с разной ценой: файл со всеми описаниями для вышестоящего портала, выборка строк таблицы для разработчика, поиск по каталогу для программы. Как они устроены — в главе «API и массовая выгрузка».

Шаг 4. Сначала то, без чего каталог невидим

Минимум одинаковый для всех, от национального портала до страницы исследовательской группы:

- постоянный адрес набора и файла;
- лицензия у каждого набора, а не только у сайта;
- дата обновления и контакт ответственного;
- выгрузка всех описаний одним файлом или запросом (небольшому сайту хватит одного файла `data.json`);
- разметка schema.org и карта сайта;
- описания читаются без регистрации и капчи.

Эти функции обязательны без отдельного обоснования, причины записаны в карте возможностей. Поиск с фильтрами, предпросмотр и статистика идут после минимума. Красивый сайт без выгрузки описаний не попадает ни в национальный портал, ни в поисковик данных.

Минимум не требует портала: статический сайт или Git-репозиторий закрывает его целиком.

Шаг 5. Расставьте уровни

Уровень	Тест	Пример обоснования
Обязательно	без функции конкретная задача не выполняется в день запуска, и понятно, кто это заметит. Внешнее обязательство всегда даёт «обязательно» — со ссылкой на документ	«Национальный портал забирает DCAT-AP, иначе наши наборы не попадут в общую витрину»
Желательно	задача выполняется и без функции, но хуже, дольше или дороже	«Предпросмотр таблицы уменьшит письма „что в файле“, но файл и так скачивается»
Не нужно	задачи нет, или её закрывает кто-то другой	«DOI не нужен: данные оперативные, на конкретную версию не ссылаются»

Пометка «не нужно» не значит «никогда». К ней пишут условие пересмотра: «DOI не нужен, пока данные оперативные; вернуться, если начнём выпускать проверенный годовой архив для исследователей». Без условия отказ через год придётся объяснять заново, а с ним видно, когда объяснение устарело.

Если обязательных функций больше половины списка, пройдите тест заново: скорее всего, часть записана «на всякий случай». По такому списку платформы не сравнить — все окажутся «почти подходящими», а цена проекта вырастет за счёт функций, которыми никто не пользуется.

Шаг 6. Оцените цену функции

Одна и та же функция стоит по-разному в зависимости от того, где она живёт.

Где функция	Цена запуска	Цена сопровождения	Главный риск
В ядре платформы	настройка	обновляется вместе с платформой	функцию меняют или убирают в новой версии, но об этом пишут в примечаниях к выпуску
В расширении	установка и настройка	проверка при каждом обновлении ядра	расширение отстаёт от ядра или его бросают
В облачном сервисе	включить в настройках или купить тариф	абонентская плата	функция зависит от тарифа и решений поставщика
Своя доработка	работа разработчика	всё сопровождение на вас	автор уходит, функция остаётся без хозяина
За пределами портала	договор или правила работы	следить, чтобы связь не порвалась	чужой сервис меняет условия или закрывается

Разница видна на живых каталогах. По реестру dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) описания в DCAT отдают 239 из 245 порталов на Socrata, где эта функция в ядре, 633 из 1906 порталов на CKAN, где её добавляет расширение ckanext-dcat, и 24 из 1144 самописных порталов открытых данных. Функция из ядра есть почти у всех установок, из расширения — у трети, из своей разработки — у 2%.

Расширение, без которого обязательная функция не работает, проверяют до выбора: кто его ведёт, на каких версиях ядра идут тесты, когда был последний выпуск. Методика — в главе «Экосистемы расширений». Пример риска: документация ckanext-doi называет только CKAN 2.9 и 2.10, а на 2.11 и 2.12 расширение не тестировалось. Если DOI обязателен, это довод искать платформу, где он в ядре.

Цену запуска и цену сопровождения оценивают отдельно. Облачный сервис запускается за недели, но платить за него придётся каждый год; свою доработку пишут один раз, а сопровождают всё время жизни каталога. Подробнее — в главе «Скорость развёртывания и сопровождение».

Функцию не обязательно делать в своём каталоге

Часть функций выгоднее отдать тому, кто их уже умеет. Тогда в списке функция остаётся, а в графе «где живёт» стоит чужой сервис.

Функция	Кто может закрыть вместо вас	Что остаётся вам
DOI и версии	Zenodo, университетский репозиторий, репозиторий консорциума	ссылка из карточки набора на DOI и правило, что считать новой версией
Поиск для жителей всей страны	национальный или региональный портал, который собирает ваши описания	выгрузка описаний в его профиле
Слои карты и картографические сервисы	картографический сервер или региональный геопортал	карточка набора, которая ведёт на сервис
История изменений небольших наборов	Git-репозиторий на GitHub или GitLab	выпуски с метками и файл `CITATION.cff`

Функция	Кто может закрыть вместо вас	Что остаётся вам
Хранение больших файлов	облачное хранилище или общий хаб данных	постоянный адрес файла, а не временная ссылка

Цена такого решения — зависимость от чужих правил: лимитов Zenodo, расписания сбора у регионального портала, тарифов облачного хранилища. Её записывают в карточку решения так же, как цену расширения. Если чужие сервисы закрывают всё, кроме минимума, свой портал, возможно, не нужен — варианты разобраны в части «Публикация без портала».

Шаг 7. Не закрывайте себе следующий шаг

Спросите, что понадобится через три-пять лет, и проверьте, какое свойство платформы сохраняет к этому дорогу:

- больше публикаторов — организации и роли в ядре, а не одна учётная запись на всех;
- второй язык — многоязычные поля описания, а не только перевод интерфейса;
- сбор муниципальных каталогов — сбор по DCAT или API в ядре или в живом расширении;
- DOI для научных приложений — связь с научным репозиторием или с DataCite;
- смена платформы — полная выгрузка описаний и файлов и адреса, которые можно перенаправить. Об этом раздел «Сколько стоит уйти».

Отказ от функции сегодня допустим. Недопустим выбор платформы, с которой к этой функции потом нельзя подойти. Облачный сервис даст свои поля в шаблоне, но не свою программу вокруг них: проверку по внешнему справочнику, выгрузку в профиль, которого нет у поставщика, новый тип записи. CKAN обрастёт и этим, но ценой расширений.

Что получается на выходе

Таблица, а не абзац «хотим современный портал». Пример ниже условный, но собран из типичных задач.

Природоохранное ведомство региона публикует ежечасные замеры станций качества воздуха, разрешения на выбросы и границы охраняемых территорий. Слои уже есть на региональном геопортале. Национальный портал собирает описания в профиле DCAT-AP. Данные готовят три отдела, публикует пока один сотрудник.

Функция	Откуда потребность	Уровень	Почему так	Как проверим	Когда пересмотреть
Постоянный адрес набора и файла	ссылки из докладов, новостей и приложений	обязательно	минимум	адрес не меняется после переименования набора и правки описания	—
Лицензия у набора	все пользователи	обязательно	без неё данные нельзя законно использовать	у двадцати случайных наборов лицензия есть и в карточке, и в выгрузке	—
Выгрузка DCAT-AP	национальный портал	обязательно	иначе наборы ведомства не попадут в общую витрину	файл выгрузки проходит валидатор профиля, национальный портал забрал описания	национальный портал сменит профиль

Функция	Откуда потребность	Уровень	Почему так	Как проверим	Когда пересмотреть
Замеры воздуха файлом и API с обновлением каждый час	разработчики приложений о качестве воздуха	обязательно	без этого приложения разбирают страницы чужих сайтов и ошибаются	данные по адресу не старше двух часов, описание полей опубликовано	—
Слои охраняемых территорий у себя	уже на региональном геопортале	не нужно	слои живут в картографическом сервере, копия разойдётся с оригиналом	карточка набора ведёт на сервис WFS и на страницу геопортала	геопортал перестанет обновлять слои или закроется
Роли для трёх отделов	задача редакции	желательно	на старте публикует один сотрудник по заявкам отделов	—	отделы начнут публиковать сами
Предпросмотр таблиц	письма «что в файле»	желательно	файл и так скачивается, словарь полей отвечает на те же вопросы	—	письма не убавятся после публикации словаря полей
DOI	никто не ссылается на версию	не нужно	замеры оперативные, на конкретное состояние не ссылаются	—	ведомство начнёт выпускать проверенный годовой архив для исследователей
Сбор муниципальных каталогов	—	не нужно	это работа регионального портала открытых данных, а не ведомства	—	регион поручит ведомству собирать экологические данные муниципалитетов

Дальше эта таблица превращается в выбор платформы: следующая глава добавляет к ней колонки «ядро, расширение, доработка, невозможно» для каждой платформы из короткого списка. Отказ от DOI из этой таблицы оформлен карточкой решения. Остальные развилки — платформа, расширение, своя разработка — оформляют так же. Четыре полные цепочки от задачи до платформы — в разобранных примерах.

Как проверить свой список

- [ ] У каждой функции записан источник: чья задача или какой документ.
- [ ] Минимум отмечен обязательным целиком.
- [ ] В списке нет решений вместо задач: названий платформ, «как у соседей», «современный».
- [ ] Каждое требование — одна функция, а не три в одной строке.
- [ ] У каждой обязательной функции есть способ проверки на живом сайте.
- [ ] Обязательных функций меньше половины списка, или про каждую понятно, почему без неё нельзя.
- [ ] Отказы записаны с причиной и условием пересмотра.
- [ ] Отмечено, какие функции закрывает не ваш каталог, и кто именно.
- [ ] В списке есть задачи программ и редакции, а не только людей на сайте.

Те же пункты, разложенные по разделам технического задания, — в чек-листе требований.

Типичные ошибки

- **Список из брошюры платформы.** Брошюра перечисляет то, что у платформы есть, а не то, что нужно вам. Сравнение по такому списку выигрывает платформа, чья брошюра.
- **Список из чужого задания.** Национальному порталу нужны сбор чужих каталогов и роли для сотни ведомств, городу с тридцатью наборами — нет. Скопированное задание переносит чужие задачи и чужую цену.
- **Функции под отчёт.** Число наборов и просмотров — показатель владельца, а не задача пользователя. Когда под него подбирают функции, получается дробление данных и витрина без выгрузки.
- **Забытая редакция.** Портал выбрали по витрине, а заводить пятьсот описаний вручную некому. Нужны массовая загрузка описаний или сбор из учётной системы, и это функция, а не мелочь внедрения.
- **«API» без содержимого.** Требование «должен быть API» выполняет любой интерфейс, даже бесполезный. Пишут, что и кому он отдаёт: все описания, строки таблицы, результаты поиска.
- **Отказ без условия.** «DOI не нужен» без условия пересмотра через год выглядит как недосмотр, а не как решение.

Источники

- Карта возможностей — минимум, функции по типам каталогов, внешние причины обязательности.
- Зачем создают каталоги и кто ими пользуется — группы пользователей и их ожидания.
- Чек-лист требований
- ISO/IEC/IEEE 29148:2018 — Systems and software engineering — Life cycle processes — Requirements engineering: свойства хорошего требования (необходимое, единичное, проверяемое и другие), требование описывает «что», а не «как».
- Доли каталогов с DCAT по платформам — выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. Запросы — в приложении «Как работать с реестром».
- Совместимость ckanext-doi — глава «CKAN: ядро и расширения».
- И. Бегтин, портал открытых данных Казахстана и общедоступные API органов власти, Telegram, 19.02.2023; отмена регистрации и лимит записей на data.egov.kz, Telegram, 26.04.2023; документация API «Талдау», проверено 12 октября 2026 года.

От функций к платформе

Короткий список из двух–четырёх платформ сравнивают по обязательным функциям. Платформа, у которой обязательная функция невозможна, выбывает до разговора о цене и о числе галочек.

Список функций к этому моменту уже есть: его собирают в главе «Принципы выбора набора функций». Здесь его прикладывают к платформам.

Шаг 1. Способ, потом марка

Сначала не CKAN и не Dataverse, а способ: чужой каталог, статический сайт, облачный сервис, свободное ПО, своя разработка. Критерии и цена ухода — в главе «Свои силы, свободное ПО или облачный сервис». Если данных мало и команды нет, сравнение CKAN с Dataverse не нужно: публикуйте в существующий каталог.

Шаг 2. Короткий список

Две–четыре платформы того класса, который совпал с типом данных и способом. Ориентир — классы ПО и карточки.

Что публикуете	Обычный короткий список
Наборы органа власти, есть своя команда	CKAN, DKAN, udata
Наборы органа власти, команды нет	Huwise, Socrata, публикация в региональный или национальный портал
Научные данные с DOI	Dataverse, InvenioRDM, публикация в Zenodo
Слои и сервисы карт	GeoNetwork, ArcGIS Hub
Показатели	PxWeb, .Stat Suite, OpenSDG
Микроданные обследований	NADA, Dataverse
Десяток файлов без команды	статический сайт или Git-репозиторий

Платформа из соседней строки в этот список не попадает. Научный репозиторий не закрывает ежедневный портал открытых данных, а CKAN с непроверенным расширением DOI не заменяет Dataverse.

Шаг 3. Матрица

По каждой обязательной и желательной функции поставьте одну из четырёх пометок:

- **ядро** — есть в самой платформе;
- **расширение** — есть в названном модуле, и вы проверили, кто его сопровождает и с какой версией ядра он работает;
- **доработка** — в платформе есть куда это встроить, объём оценивает разработчик;
- **невозможно** — платформа так не умеет и не расширяется.

Платформа, у которой обязательная функция «невозможно», выбывает. Остальных сравнивают по цене запуска и сопровождения, а не по числу галочек.

«Доработка» и «невозможно» относятся только к этому сравнению. Сводка по платформам в приложении пользуется другими словами: ядро, расширение, нет, не проверено. Клетку «нет» в матрицу переносят как «невозможно», только если дописать функцию нельзя. Клетку «не проверено» в матрицу не переносят: её закрывают по документации до того, как платформа попадёт в короткий список.

Пример для региона, которому нужны выгрузка DCAT для национального портала, сбор каталогов муниципалитетов и вход сотрудников через учётные записи организации. DOI региону не обязателен.

Функция	CKAN	Huwise	Socrata	Свой сайт на генераторе
Выгрузка DCAT	расширение ckanext-dcat, по умолчанию DCAT-AP 3	ядро; национальный профиль бывает не у всех установок, его проверяют	ядро: файл `data.json` по DCAT-US, не DCAT-AP	доработка шаблона
API поиска	ядро	ядро	ядро	невозможно без отдельной программы
Свои поля под закон	расширение ckanext-scheming	ядро (шаблоны полей)	ядро (наборы полей)	доработка
Сбор каталогов муниципалитетов	расширение ckanext-harvest	ядро: сборщики CKAN, Socrata, ArcGIS, `data.json` и других порталов Huwise; CSW включают по запросу в поддержку	ядро: Catalog Connector забирает `data.json` и публичный ArcGIS Server	невозможно
Вход через учётные записи организации	расширение (SAML или OpenID Connect)	ядро: SAML и OpenID Connect	ядро: OpenID Connect; SAML через поддержку	нет
DOI	расширение; на CKAN 2.11 и 2.12 не тестировалось	нет	нет	нет

Свои поля облачный сервис заводит в шаблонах без программиста, но в выгрузку DCAT они попадают, только если поставщик предусмотрел соответствие. У Socrata в файл `data.json` уходят поля с заранее известными именами.

Если национальный портал ждёт DCAT-AP, Socrata из этого списка выбывает: её штатный файл — DCAT-US. Huwise остаётся, пока на пилоте не видно нужного профиля. CKAN остаётся, если ckanext-dcat и расширение входа проверены на той версии ядра, которую собираются ставить. Если DOI вдруг обязателен, из этой четвёрки не остаётся никого: берут научный репозиторий, а не «CKAN плюс неизвестное расширение».

Шаг 4. Проверка на живых данных

Поставьте пилот или откройте демонстрационный стенд и опубликуйте пять–десять настоящих наборов, включая неудобный: без лицензии в источнике, с большим файлом, с ежедневным обновлением. Затем проверьте так, как написано в главах о возможностях.

- Выгрузка DCAT открывается без входа и проходит валидатор того профиля, который ждёт верхний портал.
- На странице набора есть разметка schema.org, если набор должны находить поисковики.

- Адрес набора не меняется, когда поправили заголовок.
- Файл скачивается по прямой ссылке.
- Если в матрице стоит «расширение», на стенде стоит именно оно, и видна его версия.

Пилот на демонстрационных данных поставщика эту проверку не заменяет. У облачного сервиса в пилот входит и вопрос ухода: все описания и файлы выгружаются целиком.

Когда оправдана своя разработка

Одновременно верны все четыре утверждения:

- ни одна платформа из короткого списка не закрывает обязательную функцию даже живым расширением;
- эта функция не отменяется, если сузить задачу (например, публиковать научные данные в Dataverse, а оперативные — на портале);
- есть команда не только на запуск, но и на пять лет сопровождения;
- в своей системе всё равно будут постоянные адреса, лицензия на набор, выгрузка описаний и schema.org.

Если четвёртое «сделаем потом», своя разработка повторит самую частую беду самописных каталогов: сайтом пользуются люди, а программы его не видят. По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) среди каталогов показателей, где самописных больше всего, API есть лишь у 37%.

Шаг 5. Записать решение

Каждый развилочный выбор — платформа, расширение, отказ от функции, своя разработка — оформляют карточкой решения. Четыре полных примера такой цепочки — в главе «Разобранные примеры».

Типичные ошибки

- **Восемь платформ в одной таблице.** Короткий список длиннее четырёх сравнивать уже нечем: отличия тонут. Сначала отсекают класс, потом сравнивают марки.
- **Клетка «не проверено» переписана как «нет».** Так в прежней версии этой главы сбор у Huwise стоял непроверенным, хотя в документации сервиса сборщики есть. Незакрытую клетку закрывают по документации, а не вычёркивают платформу.
- **Расширение записано как недостаток.** ckanext-dcat — обычная часть взрослого портала на CKAN. Недостатком оно становится, когда на нужной версии ядра его не тестировали.
- **Пилот на данных поставщика.** Десять чужих демонстрационных наборов не показывают ваш большой файл, ваш профиль DCAT и ваш вход.

Источники

- Сравнительные таблицы
- CKAN: ядро и расширения

- Сборщики Huwise, проверено 14 октября 2026 года
- Catalog Connector у Socrata (Tyler Data & Insights)
- Доля API у каталогов показателей — выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года.

Как обосновать решение

Каталог живёт дольше, чем люди, которые его выбирали. Карточка решения — одна-две страницы об одной развилке: что было вокруг, какие варианты правда рассматривали, по каким критериям сравнивали, что выбрали, чем выбранное лучше остальных, чем придётся платить и при каком событии выбор открывают заново. По ней следующий сотрудник видит, какое обязательство закрывает платформа.

Без карточки у него два слепых хода. Первый — не трогать платформу, потому что непонятно, что сломается. Второй — заменить её и снять обязательство, о котором он не знал: выгрузку, которую забирает национальный портал, или вход, через который публикуют сотрудники.

Коротко:

- Карточку пишут на развилку, а не на каждый пункт технического задания и не на цвет темы.
- Требование говорит, что нужно. Карточка говорит, какой способ выбрали и почему отвергли остальные.
- В разделе «почему» каждый отвергнутый вариант проигрывает по названному критерию. Слова «популярный» и «лучший» такими критериями не являются.
- Отказ от функции, которую уже просили, записывают той же карточкой, что и выбор платформы.
- «Когда пересмотреть» — событие, которое организация заметит. Дата в календаре таким событием не считается.
- Пустая клетка «почему» при заполненных остальных значит, что решение ещё не принято.
- Заменённую карточку не удаляют: помечают «заменена» и ссылаются на новую.

Требование и карточка

В главе о выборе функций список задач превращается в требования. Хорошее требование говорит, что нужно, и его можно проверить: «национальный портал забирает описания в профиле DCAT-AP». Название платформы и расширения в требовании — уже решение, его принимают следующим шагом.

Карточка — этот шаг. На то же требование она отвечает: «ставим ckanext-dcat с профилем DCAT-AP 3, свой экспорт не пишем». Полная карточка про выгрузку — в главе об экспорте метаданных. Цепочки от задачи до платформы — в разобранных примерах.

Одно требование может дать несколько карточек: платформу, расширение, отказ от соседней функции. Одна карточка может закрыть несколько требований, если их закрывает один выбор. То, чем кончилась одна карточка, часто становится контекстом следующей: выбрали CKAN — и уже отдельно решают, каким расширением закрыть вход.

На какие развилки её пишут

Развилка	Зачем карточка	Где разобрано
Способ создания: свой портал, чужой каталог, статический сайт, своя разработка	способов несколько, и ошибка дорогая	муниципалитет, ведомство с двадцатью наборами, квартальная таблица, главный адрес на своей странице, массив снимков, каталог, который ведут двое
Платформа	в коротком списке больше одной	регион и университет
Расширение, без которого обязательная функция не работает	в ядре этой функции нет	выгрузка DCAT
Отказ от функции, которую уже просили	иначе через год её поставят как забытую	отказ от DOI, город копирует большой портал
Обновление ядра, когда расширение на новую версию не проверено	на кону вход публикаторов и исправления безопасности	обновление ядра
Своя разработка	ни одна готовая платформа не закрывает обязательную функцию	свой маркетплейс
Два каталога вместо одного	две задачи плохо живут в одной программе	статслужба, регион с картами
Какой уровень качества чинить первым	иначе редакция заполняет карточки пустых файлов	«поднять качество»
Как публиковать: форма, скрипт или файл для сборщика	каталог отстаёт от системы, в которой данные уже есть	наборы уже лежат в системе
Какую лицензию писать в поле	свой текст и известная лицензия спорят, или закрытый файл помечен как открытый	свой текст или известная лицензия

Мелочи карточки не требуют: тема, цвета, подписи кнопок. Не требуют и пункты минимума, против которых никто не спорил, — постоянный адрес и сама привычка ставить лицензию на набор. Карточка появляется, когда кто-то предложил обойтись без них или заменить известную лицензию своим текстом.

Городу часто хватает одной карточки. Региону нужны карточки на платформу, на сборку расширений и на отказы. Двадцать карточек до запуска значат, что в них попали настройки.

Шаблон

Приём пришёл из разработки программ. В 2011 году Майкл Найгард описал короткую запись решения: контекст, само решение, последствия — в том числе неприятные — и пометку, если решение потом заменили. Запись пишут для человека, который придёт позже, на одну-две страницы, и хранят рядом с тем, к чему она относится. Заменённую не стирают.

Для каталога к этому добавлены сравнение и условие пересмотра. Сравнение нужно, потому что выбирают между продуктами: без отвергнутых вариантов разделу «почему» не к чему приложить. Условие пересмотра называют заранее, потому что повод обычно приходит извне — сменится профиль страны, грант или версия ядра, — и ждать, пока это заметят задним числом, уже поздно.

Раздел	Что писать
Контекст	факты, которые тянут в разные стороны: задача, ограничение, уже стоящая система
Варианты	два-четыре варианта, которые правда рассматривали
Критерии	по чему сравнивали и какой критерий важнее, если они спорят
Решение	что выбрано, одной фразой, с именами платформы и расширений
Почему	по какому критерию проиграл каждый отвергнутый вариант
Последствия	что получаем и чем платим
Когда пересмотреть	событие, после которого карточку открывают заново

В главах о возможностях карточки иногда короче: плату там уже назвал окружающий текст. Карточка, которая лежит рядом с договором, включает все семь разделов.

Как звучит раздел, который можно проверить

Один и тот же выбор — выгрузка для национального портала. Слева формулировка, по которой ничего не проверить. Справа — та, которую можно сверить с документом, валидатором или репозиторием расширения.

Раздел	Так обычно пишут	Так можно проверить
Контекст	Хотим современный портал	Регион уже на CKAN. Национальный портал забирает описания в DCAT-AP. Своей команды нет, есть подрядчик
Варианты	Рассмотреть лучшие решения	ckanext-dcat; свой экспорт силами подрядчика; ручная выгрузка раз в квартал
Критерии	Популярность, удобство, цена	Файл совпадает с профилем, который ждёт национальный портал. Кто чинит экспорт при смене профиля. Можно ли обновить CKAN
Решение	Надёжная открытая платформа	ckanext-dcat с профилем DCAT-AP
Почему	CKAN популярный	Расширение ведёт сообщество CKAN и уже умеет этот профиль. Свой экспорт станет ещё одной программой на балансе подрядчика. Ручная выгрузка отстанет в первый же месяц
Последствия	Портал станет лучше	При каждом обновлении CKAN подрядчик проверяет расширение. В ядре CKAN этой выгрузки нет, и это штатная сборка
Когда пересмотреть	Через пять лет или при необходимости	Национальный портал сменит профиль, или расширение перестанут выпускать под новую версию CKAN

Популярность иногда стоит за настоящим критерием: «в стране есть подрядчики, которые уже сопровождают текущую версию CKAN и ckanext-dcat». Тогда в карточку попадает эта фраза. Слово «популярный» само по себе сверить не с чем.

Контекст пишут фактами, без оценки. «Подрядчик может написать экспорт, и экспорт останется его программой» — факт. «Свой экспорт — плохая идея» — вывод, ему место в разделе «почему».

Варианты называют конкретно: имя расширения, чужой каталог, статический сайт с файлом `data.json`. «Более простой путь» вариантом не является. Слабый ход, которого никто не предлагал и который вставлен, чтобы победитель выглядел очевидным, портит карточку: следующий читатель не отличит его от настоящего отказа. Если смотрели только один вариант, так и пишут. Карточка тогда показывает, что сравнения не было, и это тоже полезная запись.

Критериев хватает трёх. Каждый можно проверить: файл проходит валидатор профиля, назван подрядчик, в документации расширения указаны версии ядра. Если критерии спорят, записывают, какой не отдают. Для региона совпадение с профилем страны важнее красивой витрины.

Решение — одна фраза с именами. «CKAN с ckanext-dcat» — решение. «Современная открытая платформа» — пожелание.

Раздел «почему» сравнивает. На каждый отвергнутый вариант приходится фраза, по какому критерию он проиграл. Похвала выбранному без этой фразы повторяет решение другими словами. Пустая клетка при заполненных остальных значит, что выбор ещё не сделан.

В последствиях называют плату: расширение, которое проверяют при каждом обновлении; вторая система у библиотеки; зависимость города от регионального портала; отдельного домена с витриной не будет. Раздел, в котором только выгоды, рекламирует решение. Сюда же пишут, что функции нет в ядре. Иначе следующий сотрудник примет расширение за недоделку.

«Когда пересмотреть» — событие, которое заметят без повторного чтения обоснования. Национальный портал сменил профиль. Расширение не вышло под новую версию ядра. Наборов стало больше числа, которое назвали заранее. Появилась ставка на сопровождение. Грант потребовал DOI. Рядом одной фразой — что тогда делают, если это уже ясно: «закрываем свой CKAN и переносим адреса». Если ход ещё не ясен, карточка говорит открыть вопрос заново. Новое решение — новая карточка, а не правка задним числом.

«Через пять лет», «при необходимости», «когда изменится технология» и «на следующей закупке» событием не являются: в нужный день их никто не заметит. Календарная дата уместна, когда она стоит в чужом документе, и тогда в карточке назван этот документ.

Пример отказа от DOI

Это карточка к строке «DOI — не нужно» из таблицы требований природоохранного ведомства. Образец записи, не ответ для любого портала.

Контекст. Ведомство публикует ежечасные замеры качества воздуха и разрешения на выбросы. Границы охраняемых территорий уже лежат на региональном геопортале. Национальный портал забирает описания в DCAT-AP. Исследователь просит DOI, «как у научных порталов». Договора с DataCite нет. Портал предполагают ставить на CKAN.

Варианты. Выдавать DOI на каждый набор расширением ckanext-doi. Выдавать DOI только проверенному годовому архиву, если его начнут публиковать, через Zenodo или библиотеку. Обойтись постоянным адресом набора и файла.

Критерии. Совпадает ли обещание ссылки с тем, как меняются данные. Проверено ли расширение на той версии CKAN, которую поставят. Нужен ли ведомству договор с DataCite.

Решение. DOI не выдавать. У набора и файла — постоянный адрес и дата обновления.

Почему. Замеры меняются каждый час. DOI обещает зафиксированное состояние, а открывал бы сегодняшние данные. ckanext-doi на CKAN 2.11 и 2.12 не тестировался, это записано в главе про CKAN. Договор с DataCite ради оперативных рядов ведомству не нужен. Годовой

архив для статей — другая публикация, в научном репозитории, и на неё будет отдельная карточка, когда архив появится.

Последствия. Кнопки DOI на портале нет. Письмо исследователю закрывают постоянным адресом и датой выгрузки. В ответе прямо сказано, что DOI этот портал не выдаёт.

Когда пересмотреть. Ведомство начнёт выпускать проверенный годовой архив, на который ссылаются в статьях. Тогда решают, где у архива DOI, и не включают его на этом портале молча.

Пример обновления ядра

Карточка к ситуации из главы про CKAN: новая версия ядра вышла, а расширение, без которого портал не пускает публикаторов, на ней не проверено. Факты о расширениях сверены 13 октября 2026 года. Это образец записи, не предписание всем министерствам.

Контекст. Портал министерства на CKAN 2.11. Сотрудники входят ведомственной учётной записью через ckanext-saml2auth. Последний выпуск этого расширения — 1.3.0 от 14 декабря 2022 года, в нём заявлена поддержка CKAN 2.10. Автоматические тесты идут на 2.10 и 2.11, на 2.12 их нет. CKAN 2.12 вышел 26 августа 2026 года, и проект называет его последним выпуском линейки 2.x. Версия 2.11 пока получает исправления ошибок и безопасности. Когда поддерживаемой парой станут более новые выпуски, 2.11 этих исправлений лишится. Подрядчик предлагает обновиться.

Варианты. Остаться на 2.11, пока ckanext-saml2auth не проверят на 2.12. Перейти на 2.12, заранее проверив вход на тестовом стенде. Сменить расширение на ckanext-saml: выпуск 0.4.1 от 2 июня 2026 года содержит исправление под изменения SQLAlchemy в CKAN 2.12, автоматических тестов у расширения нет.

Критерии. В день перехода публикаторы входят. Портал не остаётся без исправлений безопасности на неопределённый срок. Проверка идёт на стенде, не на боевом сайте.

Решение. Подрядчик поднимает стенд CKAN 2.12 с текущим ckanext-saml2auth и проверяет вход сотрудника. Если вход работает, переносят боевой портал. Если нет, на том же стенде пробуют ckanext-saml 0.4.1. Пока ни один вариант вход не проходит, боевой портал остаётся на 2.11.

Почему. Строка в журнале изменений не заменяет вход сотрудника: ckanext-saml поправил одну несовместимость и не показывает тестов на 2.12. Остаться на 2.11 без срока значит пропустить исправления безопасности, когда 2.11 выпадет из поддерживаемой пары. Менять расширение в день обновления, без стенда, оставляет портал без публикаторов.

Последствия. В работу по обновлению заложен стенд. В карточке записан адрес репозитория, который ставят: у ckanext-saml и ckanext-saml2auth разные команды и разная доказательная база. Если оба на стенде не пускают, обновление не откладывают молча: сразу открывают новую карточку о том, чем заменить вход.

Когда пересмотреть. Вход на стенде прошёл — эту карточку помечают заменённой, в новой записан переход и какое расширение осталось. Оба расширения не пускают — новая карточка про замену входа и про то, сколько ещё можно жить на 2.11. Вышла версия, после которой

2.11 больше не получает исправления безопасности, — оставаться на ней нельзя, даже если стенд ещё не готов: портал на это время закрывают на запись.

Если расширение перестанут выпускать

Эту развилку записывают в тот день, когда расширение ставят, а не в тот, когда оно замолчит. В разделе «когда пересмотреть» или в последствиях называют один из трёх ходов:

- остаться на последней версии ядра, с которой расширение проверено, и написать, до какого события это терпимо — обычно пока эта версия ядра получает исправления безопасности;
- заменить расширение на другое, заранее названное, и проверить замену на стенде до перехода;
- снять функцию, если она была желательной, и записать отказ отдельной карточкой.

Ход выбирают по тому, обязательна ли функция. Выгрузку, которую забирает национальный портал, снять нельзя: ищут замену или остаются на старом ядре, пока это безопасно. Кнопку цитирования снять можно: постоянный адрес и дата обращения остаются. Как оценивать само расширение — в главе об экосистемах.

Где лежит карточка и кто её пишет

Карточку пишет тот, кто выбирает, до подписания договора или до установки. Запись после запуска обычно защищает уже сделанное.

Хранят рядом с техническим заданием, в той же папке. Если конфигурация портала лежит в репозитории, карточки кладут туда же. Один файл на решение, номер по порядку, номера не используют заново. В начале — дата, статус и роль автора: «принята, октябрь 2026, ответственный за портал». Статуса три: предложена, принята, заменена. У заменённой указан номер карточки, которая её сменила.

Комитет для этого не нужен. Нужна одна страница, которую откроет человек через три года.

Как проверить свою карточку

- [ ] На карточке одна развилка.
- [ ] В контексте есть обязательство или задача и ограничение: команда, подрядчик, уже стоящая система.
- [ ] Вариантов два-четыре, каждый назван конкретно, и это те варианты, которые рассматривали.
- [ ] Критерии можно проверить, и сказано, какой важнее, если они спорят.
- [ ] Раздел «почему» говорит, по какому критерию проиграл каждый отвергнутый вариант.
- [ ] В последствиях есть плата, а не только выгода.
- [ ] «Когда пересмотреть» — событие. По нему понятно, что делать дальше, либо сказано, что откроют новую карточку.

- [ ] Отказ от функции, которую уже просили, записан отдельно.
- [ ] Для расширения, без которого обязательная функция не работает, записано, что делать, если его перестанут выпускать.
- [ ] У карточки есть дата, номер и статус. Заменённая остаётся в папке.

Типичные ошибки

- **Одна карточка на всю закупку.** Платформа, цвет шапки, DOI и подрядчик в одном тексте. Через год нельзя заменить одно, не переписывая всё. Развилки разделяют.
- **Критерии из брошюры.** «Удобный, современный, масштабируемый» выполняет любой участник сравнения. Критерий привязан к вашей задаче: профиль, который ждёт национальный портал, вход сотрудников, кто чинит расширение.
- **Чужая карточка.** Соседи выбрали CKAN с тремя расширениями, потому что собирают сотню муниципалитетов. Городу с тридцатью наборами это чужое обязательство и чужая плата.
- **Вариант для вида.** В список вставлен заведомо слабый ход, которого никто не предлагал. Настоящие отвергнутые варианты из карточки исчезают.
- **Последствия только про пользу.** Плату обнаруживают при первом обновлении ядра.
- **Дата вместо события.** «Вернуться в 2031 году» не говорит, что изменилось. Календарь уместен как срок чужого обязательства, и тогда в карточке назван документ.
- **Карточка задним числом.** Её пишут, чтобы обосновать уже подписанный договор. Варианты подбираются под победителя.
- **Старую карточку удалили.** Новый сотрудник видит текущий портал и не видит, какое обязательство сняли при замене.

Источники

- Michael Nygard, Documenting Architecture Decisions, 15 ноября 2011 года — короткая запись решения: контекст, решение, последствия, в том числе неприятные; заменённую запись оставляют и помечают.
- Принципы выбора набора функций — чем требование отличается от решения.
- От функций к платформе, разобранные примеры, свои силы, свободное ПО или облачный сервис.
- Экспорт метаданных — карточка про выгрузку DCAT.
- CKAN: ядро и расширения — версии ядра и ckanext-doi; доступ, роли и процессы — расширения входа.
- ckanext-dcat.
- Выпуск ckanext-saml2auth 1.3.0 от 14 декабря 2022 года; журнал изменений ckanext-saml, выпуск 0.4.1 от 2 июня 2026 года. Сверено 13 октября 2026 года.
- Политика выпусков CKAN; объявление CKAN 2.12 от 26 августа 2026 года — 2.12 назван последним выпуском линейки 2.x.
- Экосистемы расширений — как оценивать расширение до того, как записать его в карточку.

Свои силы, свободное ПО или облачный сервис

Способ создания выбирают раньше марки. Платформа, которая идеально закрывает функции, бесполезна, если её некому обновлять.

Шесть способов

1. **Публиковать в существующем каталоге** — региональном, национальном, университетском или общем вроде Zenodo.
2. **Статический сайт или Git-репозиторий** — JKAN, PortalJS, OpenSDG, статический каталог STAC, репозиторий на GitHub. Подробно в части «Публикация без портала».
3. **Раздел уже существующего сайта**, если движок сайта умеет отдать описания машинно, а не только нарисовать таблицу.
4. **Облачный сервис** — Huwise (бывшая OpenDataSoft), Socrata, ArcGIS Hub, Figshare.
5. **Свободная платформа у себя**, с теми расширениями, которые закрывают обязательные функции: CKAN, Dataverse, GeoNetwork, NADA, PxWeb.
6. **Своя разработка.**

Сроки запуска и сопровождения — в главе о развёртывании.

По чему сравнивать

- Закрыты ли обязательные функции ядром или живым расширением.
- Сколько людей и денег в год, включая проверку расширений при обновлении.
- Где лежат данные и можно ли так по правилам организации.
- Что будет, если поставщик, подрядчик или единственный администратор исчезнут.

Когда какой способ уместен

Способ	Уместен, когда
Чужой каталог	данных немного, своей команды нет, подходящий каталог уже принимает ваши наборы
Статический сайт или репозиторий	десятки наборов, редкие правки, есть человек, который умеет работать с Git
Раздел сайта	каталог маленький, а сайт умеет отдать DCAT или хотя бы schema.org
Облачный сервис	нужен запуск за недели, эксплуатации своей нет, данные можно разместить у поставщика
Свободная платформа	есть команда или подрядчик на годы, нужны свои поля и независимость от тарифа
Своя разработка	обязательная функция не закрывается ни одной платформой даже расширением, и команда есть на весь срок

Сколько стоит облачный сервис

Цены облачных порталов почти никто не публикует, поэтому ориентиры приходится собирать по открытым закупкам. Что удалось найти на октябрь 2026 года:

Сервис	Что известно о цене	Где это видно
PortalJS (Datopian)	открытая версия бесплатна; облачные тарифы — 1188, 3588 и 14 990 долларов в год, крупные проекты по договорённости	страница тарифов вендора
ArcGIS Hub (Esri)	базовая версия входит в платные лицензии ArcGIS Online; Premium — отдельная годовая подписка, цену называет отдел продаж	страница вариантов покупки
Huwise (бывшая OpenDataSoft)	прайса нет. Территориальное объединение Гран-Пари-Сен-Уэст в 2025 году заключило договор на 600 тысяч евро на хостинг, сопровождение и услуги; британское West of England Combined Authority заплатило 42 тысячи фунтов за год работы портала в 2026–2027 годах	объявления о закупках
Socrata (Tyler Technologies)	прайса нет. Остин в 2024 году одобрил договор на платформу Data & Insights, на которой работает городской портал, на срок до пяти лет и сумму до 5 млн долларов	материалы городского совета

Суммы из закупок нельзя делить друг на друга: в договор входят разные услуги, сроки и число пользователей, а сумма часто означает потолок, а не фактический платёж. Поэтому в карточку решения записывают не «сколько стоит у соседа», а коммерческое предложение на ваш объём. Где искать сравнимые договоры: в Европейском союзе — TED, во Франции — сведения о договорах на data.gouv.fr, в Великобритании — Find a Tender, в США — повестки и решения городских советов, в России — ЕИС.

Рынок облачных порталов существует в основном там, где принято платить за подписку. В других странах чаще ставят свободное ПО или живут на бесплатной версии ArcGIS Hub.

Своя разработка: четыре вопроса до старта

Ответы пишут в карточку решения, не в переписку.

- Какую обязательную функцию не дают ни ядро, ни живое расширение? Назовите функцию, а не «нам нужно по-особенному».
- Кто сопровождает систему через пять лет, если автор уволится?
- Как в своей системе будут устроены постоянные адреса, лицензия набора, выгрузка описаний и schema.org? Без этого каталог повторяет типичную беду самописных сайтов.
- Сколько стоит уйти? Об этом раздел «Сколько стоит уйти» в конце главы.

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) самописных каталогов около 16%. Среди каталогов показателей их 44%, и API есть только у 37%. Своя разработка чаще всего проигрывает не в день запуска, а в тот год, когда её перестают развивать.

Свой движок обычно не нужен

Соблазн написать «правильный» каталог с нуля понятен: у готовых платформ всегда есть неудобства. Но движков каталогов десятки, и большинство задач они уже решили. Два примера показывают цену соблазна.

Всемирный банк и ОЭСР. У Всемирного банка несколько отдельных публичных систем с данными: портал показателей, DataBank, каталог наборов, библиотека микроданных. В 2025 году к ним добавился ещё один портал — Data360, со своим программным интерфейсом. По наблюдению на май 2026 года, данные показателей он публикует в формате SDMX, но совместимого с SDMX интерфейса не даёт. ОЭСР пошла другим путём: её Data Explorer работает на .Stat Suite — свободной платформе, которую сообща развивают статистические организации. Новая витрина ОЭСР — это новая версия общего продукта, а не ещё одна система, которую придётся поддерживать в одиночку.

Геопортал Молдовы. Национальный портал геоданных geodata.gov.md в феврале 2024 года показывал 16 карт и слоёв и был собран отдельно от готовых геоплатформ, хотя GeoNode дал бы такой же портал из коробки. В октябре 2026 года его интерфейс работает на MapStore поверх GeoServer — тех самых компонентах, из которых состоит GeoNode, только без его каталога и карточек слоёв. Каталог при этом ведётся отдельно, на национальном геопортале инфраструктуры пространственных данных. Самостоятельная сборка из готовых частей — тоже своя разработка: склеивать, обновлять и описывать слои придётся самим.

Своя разработка оправдана, когда нужна функция, которой нет ни у одной платформы и без которой каталог теряет смысл, — и ответ на первый из четырёх вопросов выше назван конкретно.

Промежуточный путь: готовое ядро и свой интерфейс

Между «взять платформу как есть» и «написать всё самим» есть устойчивый промежуточный вариант. Хранилище описаний и файлов, программный интерфейс и сбор из других каталогов берут готовыми, а сайт для людей делают свой. Такую схему называют «безголовой» (headless): у платформы отрезан её собственный интерфейс, работает только ядро.

- Портал открытых данных Торонто — сайт на WordPress, а наборы, их описания и API живут в CKAN. В 2023 году город показывал на сайте содержимое и структуру набора, оценку качества, примеры кода; доработки выложены открытым кодом. На 12 октября 2026 года в CKAN около 550 наборов.
- Indian Data Portal исследовательского института Bharti Institute of Public Policy — сайт на Next.js поверх своего CKAN. Через API каталога сайт строит фильтры, которых в CKAN нет: детализация, отрасль, частота обновления.
- Национальный портал Канады open.canada.ca работает на CKAN с большим набором собственных расширений ckanext-canada.

PortalJS — та же идея, оформленная как продукт: готовый интерфейс поверх CKAN или другого хранилища, чтобы не писать свой. Подробнее — в главе о статических сайтах.

Плюс схемы — свой интерфейс без потери стандартного ядра: DCAT, сбор и API остаются такими же, как у тысяч других порталов. Минус — две системы вместо одной. Свой сайт должен сам держать постоянные адреса наборов и разметку schema.org, иначе поисковики увидят только ядро.

Ещё один вариант — не делать портал вовсе, а собрать национальный поисковик по уже существующим каталогам. Так поступило правительство Шотландии: вместо общего портала оно сделало find.data.gov.scot, который собирает описания из десятков местных источников. Разбор — в главе о поисковиках и агрегаторах, как устроен сбор — в главе о федерации.

Подрядчик вместо команды

Государственный портал часто делает не своя команда, а подрядчик по контракту на год. Риск такой модели в том, что вместе с контрактом кончается и развитие, а ответственность за качество размывается. Российский data.gov.ru — наглядный пример.

- В 2020 году Минэкономразвития расторгло контракт на сопровождение портала стоимостью 21,8 млн рублей: из технического задания почти ничего не было выполнено. Подрядчика внесли в реестр недобросовестных поставщиков. Так об этом со ссылкой на ЕИС писал телеграм-канал «Ах, этот Минфин» в мае 2020 года; проверить карточки в ЕИС из-за рубежа в октябре 2026 года не удалось.
- В марте 2023 года портал закрыли «на ревизию» для переноса на платформу «Гостех». Работы по контракту шли с августа 2023 года по декабрь 2024 года и стоили 30,9 млн рублей.
- Новая версия открылась только 15 июля 2025 года, с 5826 наборами против 26 490 в последнем сохранённом снимке прежней версии (февраль 2022 года). Старую версию закрыли раньше, чем была готова новая, и больше двух лет у страны не было федерального портала.

Что из этого следует для заказчика, какой бы ни была страна:

- код, описания и файлы принадлежат заказчику, и их полная выгрузка — обязательный результат каждого этапа;
- старая версия работает, пока новая не принята; если переезд неизбежен, наборы выкладывают архивом и сохраняют адреса;
- работы принимают по действующим функциям на живом сайте, а не по отчёту;
- у портала есть хотя бы один человек со стороны заказчика, который знает его устройство и переживает смену подрядчиков.

Если такого человека нет и не будет, честнее выбрать облачный сервис или готовую платформу с минимумом доработок: чем меньше своего кода, тем меньше зависимость от того, кто его писал.

Пример: свой маркетплейс данных

Самая частая идея своей разработки, которая проваливается не из-за кода, а из-за экономики.

Контекст. Организация покупает данные у нескольких поставщиков и хочет сделать площадку, где поставщики выставляют наборы, а покупатели их находят и оплачивают. Готовые свободные платформы каталогов оплату не принимают.

Варианты. Своя площадка. Облачная площадка (AWS Data Exchange, Snowflake Marketplace), если поставщики и покупатели уже в этом облаке. Открытый каталог описаний без оплаты: договор и доставка идут напрямую между сторонами.

Критерии. Откуда придут поставщики и покупатели. Чем площадка выгоднее прямого договора с поставщиком. Кто платит за продвижение, пока площадка пустая.

Решение. Открытый каталог описаний, а для продажи — облачная площадка, если она подходит.

Почему. Площадки данных живут там, где покупатели уже есть: у облачных провайдеров, у поставщиков программ для трейдеров, у государства, которое и продвигает площадку, и может быть регулятором. Независимой витрине поставщиков пришлось годами тратиться на продвижение. Владельцы ценных данных посредника не ищут, а покупка данных редкая и договорная. Своя площадка без готовой аудитории остаётся пустой.

Последствия. Нет своей системы оплаты — нет её поддержки. Описания открыты, поэтому покупатель и поисковик видят, что данные существуют.

Когда пересмотреть. Появились поставщики и покупатели, которые договорились работать через вас, и на продвижение есть бюджет.

Сколько стоит уйти

Цену ухода считают до подписания договора. Иначе её узнают в момент, когда уходить уже надо.

Сложите четыре статьи:

1. **Выгрузка.** Все описания и все файлы, не выборочно. У свободных платформ и у честного облачного сервиса есть API или дамп. Если выгрузки нет, это уже ответ.
2. **Адреса.** Таблица старых и новых адресов и перенаправления. DOI, выданные заранее, эту статью уменьшают: ссылка `doi.org` переезжает вместе с записью в DataCite.
3. **Люди и простой.** Кто переносит, сколько недель каталог открыт только на чтение, кто пишет пользователям.
4. **Повторное обучение публикаторов.** Новая форма набора стоит дороже, чем кажется из технического задания.

Ориентир для сравнения, не смета:

Откуда уходите	Что обычно дорого
Чужой каталог или Zenodo	почти нечего переносить, если своих адресов не обещали
Статический сайт	дёшево: файлы и страницы уже у вас
Облачный сервис	выгрузка и смена адресов; договор стоит прочитать до оплаты
Свободная платформа	свои расширения и своя схема полей; сами файлы забираются

Откуда уходите	Что обычно дорого
Своя разработка	всё, если автор не оставил дамп и карту адресов

Привязка к облачному сервису видна и по закупкам. Гран-Пари-Сен-Уэст в 2025 году заключил новый договор на сопровождение своего портала на OpenDataSoft без конкурса: в объявлении сказано, что права на продукт и его сопровождение есть только у вендора. Это законно и честно записано, но означает, что торговаться о цене после внедрения уже не с кем.

Закрытие чужого продукта бьёт по всем сразу. Nesstar, на котором работали многие архивы микроданных, перестали развивать в 2015 году, и каждый архив переносил коллекции сам; канадский ODESI выбрал новую платформу и начал переезд до отключения старой. Смена владельца Socrata — тот же класс риска для облака. Карточка решения на облачный сервис без абзаца «как заберём данные» неполная.

Источники

- Разобранные примеры
- Доли самописных каталогов — выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года.
- И. Бегтин, Про рынки и биржи данных, сентябрь 2026 — модели площадок данных и почему они редко окупаются
- Цены облачных сервисов, проверено 12 октября 2026 года: тарифы PortalJS; варианты покупки ArcGIS Hub; TED, извещение 867108-2025 — Grand Paris Seine Ouest и OpenDataSoft, 600 000 евро, без конкурса из-за исключительных прав; West of England Combined Authority, Open Data Portal Contract, 2026; City of Austin, Legislation Text 24-4611, 2024 — Tyler Data & Insights
- И. Бегтин, экономика порталов открытых данных, Telegram, 07.03.2026 — обзор вендоров; приведённые там цены по сайтам вендоров и закупкам не подтвердились, кроме PortalJS
- Всемирный банк и ОЭСР: Data360 и его API, DataBank, каталог наборов; OECD Data Explorer и сообщество SIS-CC, развивающее .Stat Suite; И. Бегтин, о Data360, Telegram, 15.05.2026
- Молдова: geodata.gov.md — интерфейс на MapStore и GeoServer, проверено 12 октября 2026 года; национальный геопортал ИПД; И. Бегтин, о geodata.gov.md, Telegram, 07.02.2024
- Готовое ядро и свой интерфейс: open.toronto.ca и доработки CKAN; Indian Data Portal; ckanext-canada — число наборов и состав расширений по `status_show` и `package_search`, 12 октября 2026 года; И. Бегтин, Telegram: портал Торонто, 16.11.2023; Indian Data Portal, 20.09.2024; CKAN как система управления данными и PortalJS, 26.05.2023
- data.gov.ru: «Если быть точным», Портал открытых данных возобновил работу после двухлетнего перерыва, 16.07.2025 — 30,9 млн рублей, сроки контракта, число наборов; «Коммерсант», Минэкономики готовится к перезапуску портала открытых данных data.gov.ru, июль 2025; ГАРАНТ.РУ, Портал открытых данных начал работу после масштабного обновления, 15.07.2025; «Ах, этот Минфин», поставщик портала открытых данных добавлен в РНП (репост в канале @begtin), 08.05.2020, и о контракте на перенос

портала на Гостех, август 2023 — 30,9 млн рублей при начальной цене 50,7 млн, расторжение контрактов 2019–2020 и 2020–2022 годов; И. Бегтин, о контракте на перенос портала, Telegram, 14.08.2023

Типовые сценарии

Сценарий отвечает на вопрос «с чего начать», а не «что закупить». Строка ниже называет тип каталога, функции, без которых этот владелец не выполнит свою работу, и короткий список платформ на проверку. Платформу из соседней строки не переносят: у города и у региона разные обязанности и разная команда.

Коротко:

- Сначала найдите свою строку в таблице. Она задаёт тип и обязательные функции. Платформу выбирают следующим шагом, по списку функций и матрице.
- К каждой строке добавляется минимум для любого каталога: постоянный адрес, лицензия у каждого набора, дата обновления и контакт, выгрузка описаний, разметка для поисковиков, описания без регистрации. В строке — только то, что добавляется сверх минимума.
- Если владелец делает два дела, у него две строки и часто две платформы.
- Четыре строки доведены до карточки решения в главе «Разобранные примеры»: муниципалитет, регион, университет, статистическая служба. Для своего города разделы карточки заполняют заново. Чужая карточка тащит чужое обязательство и чужую плату.

Цифры — по выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года).

Как пользоваться

1. Найдите ближайшего владельца в таблице.
2. Выпишите столбец «что добавить к минимуму». Это кандидаты в обязательные требования, не готовое техническое задание. Проверка на обязательность — в главе о выборе функций: какая задача не выполнится в день запуска без этой функции и кто это заметит.
3. Если подходят две строки, откройте раздел «Две задачи сразу».
4. Короткий список из своей строки проверьте по матрице «функция × платформа». Платформа, у которой обязательная функция невозможна, выбывает до разговора о цене.
5. Развилку — свой портал или чужой, какую платформу, какое расширение — запишите карточкой решения. Пункты для задания собраны в чек-листе.

Кто	Тип каталога	Что добавить к минимуму	Куда дальше
Центральный орган власти	портал открытых данных или сборщик чужих каталогов	профиль описаний, который ждёт страна, и постоянный адрес выгрузки	ниже
Регион	портал, который собирает города и сам отдаёт наборы стране	приём чужих описаний по расписанию и своя выгрузка в национальном профиле	пример региона
Муниципалитет	чаще чужой каталог, а не свой портал	набор виден в региональном или национальном портале	пример муниципалитета
Картографическая служба	геопортал	описания слоёв и каталожный сервис	геопорталы
Статистическая служба	два каталога: показатели и микроданные	выборка по разрезам и обмен SDMX; описание обследования и уровни доступа	пример статслужбы

Кто	Тип каталога	Что добавить к минимуму	Куда дальше
Университет и научная организация	научный репозиторий	DOI, версии, цитирование, файлы у себя	пример университета
Научный консорциум	предметный репозиторий своей дисциплины	свои поля, протокол для агрегаторов дисциплины, план долгого хранения	научные репозитории
Международная организация	хаб из каталогов разных типов	несколько языков, стабильные адреса, выгрузка для агрегаторов	гибриды
Небольшая организация, журналисты, группа	публикация без портала	разметка schema.org на странице набора	когда портал не нужен
Компания	портал наборов, каталог API или маркетплейс	лицензия, понятная юристу; описание открыто, даже если данные платные	другие типы
Лаборатория и программа данных для моделей	обычно чужой хаб, не свой каталог	карточка набора, фиксированное разбиение на части, лицензия на обучение	ML-каталоги

Центральный орган власти

Под этим названием смешивают четыре разные работы. Короткий список у них разный.

Работа	Короткий список
Все ведомства публикуют на одном сайте	CKAN, DKAN или udata: организации, роли, модерация. Сбор чужих каталогов добавляют, когда часть ведомств остаётся на своих сайтах. Так устроен data.gouv.fr: и площадка, и сборщик
Сайт почти ничего не хранит и собирает чужие каталоги	Piveau или CKAN с ckanext-dcat и ckanext-harvest. Если каталоги ведомств уже есть, страна может начать с поисковика по ним: find.data.gov.scot собирает местные источники, общий портал data.gov.scot появился позже
Ведомство отдаёт свои наборы наверх и общий портал не заменяет	CKAN с ckanext-dcat. Облачный сервис — если файл выгрузки совпадает с профилем страны. У Socrata штатный файл — DCAT-US. У Huwise национальный профиль подрядчик видит на пилоте у поставщика и в репозитории сам не правит
Страна раздаёт городам одну и ту же сборку	Сборку выбирает центр, не город: Andino, GD Catalog, государственная платформа сайтов. У каждой витрины общие поля и общая выгрузка наверх

Тип. Портал открытых данных. У сборщика это каталог описаний со ссылкой на первоисточник, свои файлы ему почти не нужны.

Обязательно. Профиль описаний, который уже читает адресат, и устойчивый адрес, откуда описания забирают. В ЕС это DCAT-AP и национальные уточнения к нему, в США — DCAT-US. В России федеральный портал и сайты на Gosweb ждут паспорт методических рекомендаций 4.0, а не DCAT. Паспорт оставляют для российского портала. Разметку schema.org и DCAT добавляют, когда наборы должен найти кто-то ещё или когда сборщик назвал профиль. Как развести эти языки — в главе о стандартах метаданных.

Кандидаты. Из таблицы выше, и только те, у кого нужный профиль есть в ядре или в живом расширении. Среди самописных порталов открытых данных выгрузку DCAT реестр нашёл у 24 из 1 144. Сайт откроют люди. Национальный портал описания не заберёт.

Подробнее. Порталы открытых данных, экспорт метаданных, CKAN и расширения.

Когда открыть карточку. Профиль страны содержит свои поля, которых нет в общем DCAT-AP. Появилась вторая полка — карты, показатели, микроданные. Страна выбирает, раздавать ли городам одну сборку.

Регион

Тип. Портал открытых данных, который и собирает муниципалитеты, и сам отдаёт описания стране.

Обязательно. Приём чужих описаний по расписанию и своя выгрузка в национальном профиле. Если в профиле есть поля страны сверх общего DCAT-AP, они попадают и в форму, и в файл, который забирает национальный портал. В карточке копии стоит ссылка на страницу набора у города, иначе один набор посчитают трижды: у города, у региона и у страны.

Кандидаты. CKAN: ядро, ckanext-dcat и ckanext-harvest. Свои поля задаёт ckanext-scheming, в выгрузку их кладёт профиль ckanext-dcat. Если страна принимает общий DCAT-AP без своих полей, scheming не ставят. Piveau ставят, когда портал в основном собирает чужие каталоги, а профиль страны уже есть в продукте: форма ввода документирует DCAT-AP и DCAT-AP.de. Профиля, которого в этом списке нет, в короткий список не кладут: его дальше сопровождает подрядчик.

Socrata и Huwise сбор умеют в самом сервисе, но на этой строке обычно выбывают по профилю. У Socrata штатный файл — DCAT-US. У Huwise свои поля попадают в DCAT, только если поставщик предусмотрел соответствие. Полное сравнение и отвергнутые варианты — в примере региона.

Подробнее. Федерации и агрегация, CKAN и расширения.

Когда открыть карточку. Карточка в примерах записана для профиля DCAT-AP с несколькими своими полями и подрядчика без своей команды разработки. Другой профиль или обязанность сидеть на одной государственной платформе — новая карточка.

Муниципалитет

Тип. Чаще всего не свой портал. В реестре около 17,6 тысячи каталогов местных органов власти, из них около 14,4 тысячи — геопорталы. Порталов открытых данных у местных органов 2 862. Заметная доля — витрины одной сборки: 421 раздел на российской платформе сайтов Gosweb, общие каталоги поставщика за городскими витринами в Японии и Италии. Платформу в таком случае выбирает не город.

Обязательно. Набор виден в региональном или национальном портале. У него есть лицензия и дата обновления. Файл описаний совпадает с профилем, который этот портал собирает. У статического каталога JKAN файл `data.json` объявляет схему DCAT-US 1.1, шаблон заполняет короткий список полей. Регион, который ждёт DCAT-AP, получит описания после того, как шаблон переделают под его профиль.

Кандидаты. Публикация организацией на региональном портале, в том числе в типовом разделе, который регион уже ведёт для городов. Для карт — геопортал, который уже ведёт регион или страна. Для десятка наборов, когда регион чужие организации не заводит, — страница на сайте города с разметкой и файлом описаний.

Подробнее. Разобранный пример муниципалитета, когда портал не нужен.

Когда открыть карточку. В примере города порог такой: наборов стало больше двухсот, появилась ставка на сопровождение портала, или регион перестал принимать муниципальные данные. До этого дня свой CKAN не ставят.

Картографическая служба

Тип. Геопортал. Под этим словом три разные части: просмотрщик показывает слои на карте, каталог описывает их, сервисы отдают слой программам. Картографический сервер без каталога — ещё не каталог. Из 4 674 геопорталов на ArcGIS Server каталожный сервис реестр нашёл у одного, из 1 927 муниципальных порталов e-mapa — ни у одного.

Обязательно. Описания слоёв по ISO 19115 и каталожный сервис: CSW или OGC API Records. У слоя есть лицензия. Слой можно скачать, или на карточке прямо сказано, что доступен только просмотр. Если слои входят в инфраструктуру пространственных данных страны, карточки отдают в её профиле. Выгрузка DCAT нужна дополнительно, когда тот же слой должен попасть на портал открытых данных.

Кандидаты. GeoNetwork — для каталога. Сервисы оставляют на том сервере, который уже стоит: GeoServer или ArcGIS Server. ArcGIS Hub — если организация уже работает в ArcGIS и ей нужны и карточки, и скачивание файлами. GeoNode — когда одна программа должна и показать карту, и вести каталог, а отдельной команды на две системы нет. Маленькому городу свой сервер обычно вреден: слои отдают в региональную или национальную инфраструктуру.

CKAN в этот список не входит. Расширение ckanext-spatial забирает описания слоёв в поиск портала открытых данных и показывает ссылку на сервисы. Каталог инфраструктуры оно не заменяет.

Подробнее. Геопорталы, карточка «регион с картами и наборами».

Когда открыть карточку. Служба хочет закрыть слои и таблицы одним сайтом, или страна требует сразу и каталог инфраструктуры, и выгрузку DCAT.

Статистическая служба

Тип. Два каталога: показатели и микроданные. Показатель выбирают по году и территории. Микроданные — обезличенные файлы обследования: описание открыто, файл по заявке. Один самописный сайт обычно плохо делает оба дела.

Обязательно. Для показателей — выборка по разрезам и обмен по SDMX со структурами и справочниками. Для микроданных — открытое описание обследования в DDI и уровни доступа к файлам.

Кандидаты. Микроданные — NADA. Показатели — .Stat Suite, если обмен по SDMX главная обязанность и есть консультанты: служба SDMX у неё в ядре. Если команда уже умеет PxWeb, витрину ставят на PxWeb, а точку SDMX добавляют отдельно. Полноценной службы SDMX у PxWeb нет: старый API отдаёт файл SDMX-ML 2.0, в PxWebApi 2 формата SDMX нет. Показатели внутри NADA не включают. С выпуска 5.6 платформа умеет их описывать, но в

библиотеке микроданных Всемирного банка все 7 145 записей — обследования, а показатели банк публикует отдельно. Nesstar для нового каталога не рассматривают: его объявили устаревшим и к 2022 году сняли.

Среди каталогов показателей самописных 44% (1 505 витрин). Программный доступ есть у 37% каталогов этого типа и у 8% самописных. Две полные карточки — в примере статистической службы.

Подробнее. Каталоги показателей, каталоги микроданных.

Когда открыть карточку. Карточки в примерах уже разводят полки и выбирают платформу показателей. Новая нужна, когда NADA перестанут выпускать или обмен уйдёт с SDMX. На октябрь 2026 года выпуск NADA 5.7.1 вышел 4 октября.

Университет и научная организация

Тип. Научный репозиторий, а не портал открытых данных с доработками. Открытые данные администрации — бюджет, расписание, адреса — публикуют на странице университета или на портале органа власти. В репозиторий данных они не попадают.

Обязательно. DOI, версии и цитирование, которое можно выгрузить в BibTeX. Файлы лежат в репозитории, а не по ссылке на сайт группы. Библиотека модерирует запись. Свой префикс DOI нужен, когда университет хочет, чтобы идентификатор был его, а не общий префикс Zenodo.

Кандидаты. Dataverse, если библиотека готова вести вторую систему рядом с уже стоящим репозиторием публикаций: DOI, версии и цитирование у Dataverse в ядре, выгрузки DCAT в ядре нет. InvenioRDM — если есть кому сопровождать Python-приложение и университет сам отдаёт описания в DCAT: выгрузка DCAT у него в ядре. Уже стоящий DSpace оставляют репозиторием публикаций. DOI в нём включается настройкой, версии есть, но до версии 10.1 включительно готовой ссылки и выгрузки в BibTeX и RIS в ядре нет. Zenodo или общий репозиторий консорциума — когда библиотеке некому вести вторую систему: DOI и цитирование там есть, своего префикса и модерации групп нет.

CKAN с расширением ckanext-doi в этот список не входит. На CKAN 2.11 и 2.12 расширение не тестировалось, версий набора в ядре нет.

Подробнее. Разобранный пример университета, идентификаторы и DOI, научные репозитории.

Когда открыть карточку. Карточка в примерах записана для библиотеки без отдельной команды на новое приложение и без обязанности отдавать DCAT на первом шаге. Появилась эта обязанность или вышла версия DSpace с выгрузкой цитирования в ядре — карточку открывают заново.

Научный консорциум

Тип. Предметный репозиторий со схемой своей дисциплины. Пример — PANGAEA для наук о Земле. NIH советует автору сначала искать репозиторий своей дисциплины, и только если его нет — универсальный или университетский. Консорциум заводит свой, когда у дисциплины

ещё нет репозитория, который принимает её формат.

Обязательно. Свои поля описания: здесь они важнее, чем у городского портала. DOI и версии. Протокол, которым данные забирают общие агрегаторы дисциплины, например OAI-PMH. Если репозиторием будут пользоваться как архивом, а не как папкой проекта, — план долгого хранения и со временем сертификация вроде CoreTrustSeal.

Кандидаты. Dataverse или InvenioRDM со своей схемой метаданных. Отраслевая система, если дисциплина уже договорилась о ней: GBIF IPT для данных о биоразнообразии, другие инструменты той же дисциплины для наблюдений Земли и климата. Своя разработка — когда ни одна из них не держит обязательный формат данных, и есть команда на годы сопровождения, а не только на запуск.

Подробнее. Научные репозитории, полнота метаданных, методики и рейтинги.

Когда открыть карточку. Дисциплина уже ведёт общий репозиторий, и консорциум решает, подключаться к нему или ставить свой. Сертификация хранения становится условием фонда.

Международная организация

Тип. Часто хаб: несколько каталогов разных типов под одной вывеской и с общим поиском. Всемирный банк держит их раздельно: показатели, библиотека микроданных на NADA и общий каталог наборов, который собирает описания из остальных. На 12 октября 2026 года в каталоге наборов было 7 900 записей. Тихоокеанское сообщество собрало Pacific Data Hub: портал наборов на CKAN, микроданные на NADA, показатели на .Stat Suite, геопорталы на GeoNode, картографическая витрина на TerriaJS, серверы THREDDS и каталог снимков STAC. Человеку, который пришёл за данными одной страны, не нужно знать, на какой полке они лежат.

Обязательно. Несколько языков интерфейса и описаний. Стабильные адреса: на данные ссылаются отчёты на годы вперёд. Выгрузка описаний для агрегаторов, и с каждой полки — на её языке: DCAT с портала наборов, SDMX с показателей, DDI с микроданных. Если организация финансирует проекты развития, данные о них публикуют по стандарту IATI: его читают доноры и агрегаторы помощи. Общий поиск хаба строят сбором описаний в один каталог. Этот сбор — ещё одна цепочка, за ней нужен человек. Хаб, который человек открывает в браузере, а программа не читает, агрегатор снаружи не соберёт.

Кандидаты. Для каждой витрины — платформа её типа: CKAN, DKAN, udata или облачный сервис для файлов, .Stat Suite или PxWeb для показателей, NADA для микроданных, GeoNetwork или GeoNode для слоёв. Один движок на все полки — обычная причина самописной системы, которую потом некому сопровождать. Рабочая схема — несколько понятных входов и общий поиск поверх них.

Подробнее. Сравнение типов и гибриды, другие возможности: несколько языков.

Когда открыть карточку. Хаб заводят, когда каталогов уже три и больше и человеку важно искать по всем сразу. До этого достаточно двух адресов на сайте организации и подписи, чем полки отличаются.

Небольшая организация, журналисты, исследовательская группа

Тип. Портал не нужен. Десяток файлов, одна команда, обновление несколько раз в год закрываются страницей и файлом описаний.

Обязательно. Постоянный адрес, лицензия, разметка schema.org, дата и контакт. Если на выпуск нужно сослаться в статье, DOI получает выпуск в общем репозитории, а не каждый файл на своём портале.

Кандидаты. Статический сайт. Git-репозиторий с выпусками в Zenodo. Общий хаб, если аудитория уже там. Свой CKAN, Dataverse или облачная подписка на этом размере добавляют сопровождение, которого некому вести.

Подробнее. Часть «Публикация без портала»: статические сайты, Git-репозитории, как сделать публикацию находимой.

Когда открыть карточку. Наборов стало столько, что файл описаний и страница поиска перестают справляться, или закон требует именно портал. Пороги по типам собраны в главе «Когда портал не нужен». Для ведомства с двадцатью наборами карточка уже записана там же.

Компания

Тип. Зависит от того, что компания открывает. Открытые наборы — небольшой портал или страница. Программные интерфейсы — каталог API для разработчиков. Продажа данных — маркетплейс. Это три разных каталога, и удачный образец одного не переносится на другой.

Обязательно. Лицензия, понятная юристу покупателя или пользователя. Описание открыто, даже если файл платный: иначе покупатель не видит, что покупает, а поисковик не находит набор.

Кандидаты. Облачный сервис или аккуратный статический сайт для открытых наборов. Портал разработчика из системы управления API — для каталога API; описание интерфейса отдают в OpenAPI. Для продажи — облачная площадка вроде AWS Data Exchange, если поставщики и покупатели уже в этом облаке, или открытый каталог описаний, а договор и доставка идут между сторонами напрямую. Свободные платформы каталогов оплату не принимают. Своя площадка без готовой аудитории обычно остаётся пустой: разбор — в карточке «свой маркетплейс».

Подробнее. Каталоги API, маркетплейсы и списки наборов.

Когда открыть карточку. Компания хочет и публиковать открытые наборы, и продавать другие, и показать разработчикам API. Это три строки, у каждой свой короткий список.

Лаборатория и данные для моделей

Тип. Обычно не свой каталог. Наборы для обучения моделей ищут на нескольких больших площадках, и туда же смотрят конференции и библиотеки. В реестре 100 ML-каталогов. На Hugging Face 12 октября 2026 года было больше миллиона наборов. Свой каталог оправдан,

когда данные нельзя отдать на чужую площадку или когда государственная программа учитывает наборы многих организаций.

Обязательно. Карточка набора: состав, способ сбора, известные перекосы, для чего набор не годится. Разбиение на обучающую и тестовую части зафиксировано и не меняется между выпусками. Лицензия отвечает, можно ли обучать на наборе коммерческую модель. Если набор сопровождает статью на конференции, которая требует описание в Croissant, карточку отдают в нём. Копию выпуска с DOI кладут в научный репозиторий: площадка может закрыться, ссылка из статьи — нет.

Кандидаты. Hugging Face или Kaggle и копия выпуска в Zenodo или в университетском репозитории. Соревнование со скрытой тестовой частью — Codabench или соревнования Kaggle. Данные, которые нельзя выложить на чужую площадку, — Dataverse, у которого Croissant в ядре с версии 6.10: описание открыто, файл по заявке. Портал открытых данных отдельный ML-каталог не заводит: таблицы отдают со словарём полей и, если платформа умеет, с описанием в Croissant. Коллекция из десятка наборов — статический сайт с файлом Croissant у каждого набора.

Подробнее. ML-каталоги, облачные хранилища и общие хабы.

Когда открыть карточку. Программа должна принимать наборы от многих организаций и не может отдать файлы на чужой хаб. До этого дня свой каталог не ставят.

Две задачи сразу

Один владелец часто узнаёт себя в двух строках. Портал без выгрузки DCAT теряет федерацию открытых данных, геопортал без каталожного сервиса — инфраструктуру геоданных, репозиторий без DOI — цитирование. Эти цепочки почти не пересекаются, поэтому платформа, удачная для одной строки, редко закрывает соседнюю.

Вместе	Как обычно устроено	Где разобрано
Регион публикует наборы и слои	портал открытых данных и геокаталог рядом; общий поиск забирает описания слоёв и ссылается на сервисы	карточка региона с картами
Служба ведёт показатели и микроданные	две платформы, на сайте две полки с разными подписями	пример статслужбы
Университет хранит статьи и данные	репозиторий публикаций и репозиторий данных; связь держат идентификаторами в описаниях	пример университета
Организация ведёт файлы, ряды и обследования	хаб: у каждой полки своя платформа, общий поиск сверху	три схемы гибрида

Вторую, маленькую задачу иногда закрывают расширением первой платформы: предпросмотр слоя на портале открытых данных, описание API рядом с наборами в udata. Обязательный стандарт второй задачи расширением не подменяют. Если инфраструктура геоданных ждёт ISO 19115 и сервисы слоёв, файлы в CKAN это обязательство не закрывают.

Если своей строки нет

Каталог только описаний — справочники, реестр репозиториев, паспорт показателя без самих чисел — это каталог метаданных. Поиск по чужим каталогам без своих наборов — поисковик или агрегатор. Тип ещё не ясен — его даёт маршрутная карта по вопросу «что главное в данных», а не по названию должности.

Источники

- Разобранные примеры, маршрутная карта, принципы выбора функций, от функций к платформе, как обосновать решение, чек-лист требований
- Карта возможностей — минимум для любого каталога и что добавляет тип
- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: около 17,6 тысячи каталогов местных органов власти, из них около 14,4 тысячи геопорталов; 2 862 портала открытых данных у местных органов; 421 витрина Gosweb; каталожный сервис у 1 из 4 674 геопорталов на ArcGIS Server и ни у одного из 1 927 порталов e-mapa; DCAT у 24 из 1 144 самописных порталов открытых данных; 1 505 самописных каталогов показателей, API у 37% каталогов типа и у 8% самописных. Запросы — в приложении «Как работать с реестром»: местные органы, порталы открытых данных, геопорталы, карта возможностей, каталоги показателей
- Порталы открытых данных — модели национального портала, типовые сборки, паспорт и Gosweb; стандарты метаданных — паспорт рекомендаций 4.0 и когда к нему добавляют DCAT
- Геопорталы, каталоги показателей, каталоги микроданных — NADA 5.7.1 от 4 октября 2026 года и снятие Nesstar, научные репозитории, ML-каталоги
- Сравнение типов и гибриды — три схемы, карточка региона с картами, состав Pacific Data Hub и каталог наборов Всемирного банка (7 900 наборов на 12 октября 2026 года, 7 145 обследований в библиотеке микроданных)
- PANGAEA; Всемирный банк: показатели, библиотека микроданных, каталог наборов; Pacific Data Hub; стандарт IATI
- find.data.gov.scot — поисковик по уже существующим каталогам; разбор — в главе о поисковиках
- И. Бегтин, Общедоступные каталоги данных международных организаций, январь 2023 — хабы из нескольких каталогов

Чек-лист требований

Пункт без ссылки на объяснение превращается в строчку, которую копируют из чужого задания. Ниже у каждого пункта есть глава, где сказано, когда без него можно обойтись. Какие пункты типичны для министерства, города, университета или компании — в главе «Типовые сценарии». Строка сценария задаёт черновик обязательности, в задании её проверяют заново.

В задании у пункта три пометки: обязательно, желательно, не нужно — и одна фраза, откуда потребность. Пустое «обязательно» на всё подряд делает сравнение платформ бессмысленным.

Минимум для любого способа публикации

Даже для статического сайта из десятка наборов.

- [ ] Постоянный адрес каждого набора и файла — идентификаторы. Обязательно всем.
- [ ] Лицензия на каждый набор, не только на сайт — лицензии. Обязательно всем.
- [ ] Файл со всеми описаниями — экспорт метаданных. Обязательно любому каталогу из нескольких наборов, в том числе небольшому сайту: ему хватит одного файла `data.json` или `datapackage.json`. Странице с одним набором не нужен.
- [ ] Карта сайта sitemap и разметка schema.org — экспорт метаданных. Обязательно, если набор должны найти поиском.
- [ ] Дата обновления и периодичность — полнота метаданных. Обязательно, если данные обновляют.
- [ ] Контакт ответственного. Обязательно всем.
- [ ] Описания читаются без регистрации и капчи — API и выгрузка. Обязательно всем, кого могут собирать автоматически.

Включать, только если есть задача

- [ ] API поиска — обязательно, когда наборов сотни и больше. Не нужно статическому сайту.
- [ ] Своё хранилище файлов и API к таблицам — хранилище. Обязательно научному репозиторию (файлы). Желательно порталу открытых данных. Не нужно геопорталу, который ссылается на сервисы.
- [ ] DOI и версии — идентификаторы. Обязательно научному репозиторию. Не нужно оперативным данным.
- [ ] Экспорт цитирования — цитирование. Обязательно научному репозиторию. Желательно, если на данные ссылаются в аналитике.
- [ ] Свои поля и несколько языков — полнота метаданных. Обязательно, если этого требует закон или страна. Иначе желательно или не нужно.

- [ ] Роли, модерация, ограниченный доступ — доступ и процессы. Роли обязательны, когда публикует больше одной команды. Ограниченный доступ обязателен для микроданных и чувствительных файлов.
- [ ] Сбор чужих каталогов — другие возможности. Обязательно региональному и национальному порталу. Не нужно ведомству, которое публикует только своё.
- [ ] Поиск по тексту описаний, предпросмотр, статистика и форма обратной связи — другие возможности. Поиск обязателен, когда наборов сотни и больше. Остальное включают по порогу из главы: есть задача и есть кому этим пользоваться.

Новые возможности

Их пока нет в ядре массовых платформ. Пункт пишут, только если задача названа. Что стоит за каждым, — в главе о передовых каталогах.

- [ ] API к каждому табличному файлу без ручной загрузки, с профилем колонок и описанием запросов. Желательно порталу с сотнями таблиц, которые фильтруют. Не нужно, если файлы скачивают целиком.
- [ ] Проверка файла по схеме при загрузке, с отказом при ошибке. Желательно, если для таблиц уже есть общие схемы. Обязательно архиву области с общим стандартом, вроде BIDS у нейровизуализации.
- [ ] Сервер MCP только на чтение, без ключа для открытых наборов, поверх публичного API, с инструкцией для агента — ИИ и пространства данных. Желательно, если его даёт платформа и лицензии с выгрузкой уже на месте. Не нужно писать свой ради отчёта.
- [ ] DOI на сохранённую выборку или выпуски с названным сроком хранения — цитирование. Желательно, если на срезы данных ссылаются в работах. Не нужно оперативным данным.
- [ ] Постоянные идентификаторы объектов внутри данных и таблица соответствия со старыми номерами — идентификаторы. Обязательно реестру и справочнику. Не нужно разовому отчёту.
- [ ] Ступенчатый доступ: описание и сводная статистика открыты, данные по ступеням. Обязательно для чувствительных данных, к которым нужен доступ исследователей. Не нужно открытым данным.
- [ ] Выборка и расчёт на сервере. Желательно, если данные не помещаются в файл. Каталог и вычисления описывают отдельными пунктами: закрытие вычислений не должно закрыть каталог.

По типу каталога

- Геопортал: каталог ISO 19115 и сервис CSW или OGC API Records; слой можно скачать или честно сказано, что доступен только просмотр.
- Научный репозиторий: DOI, версии, OAI-PMH, экспорт цитирования.
- Каталог показателей: SDMX или другой машинный доступ к таблице, описание методики.
- Каталог микроданных: уровни доступа, описание исследования, журнал заявок.

- ML-каталог: описание в Croissant, версии набора с неизменным делением на обучающую и тестовую части, лицензия и происхождение данных.

Когда сравниваете платформы

- [ ] У каждой обязательной функции пометка: ядро, какое расширение и кто его сопровождает, доработка или невозможно — от функций к платформе.
- [ ] На каждую развилку — платформу, обязательное расширение, отказ от запрошенной функции — есть карточка решения с событием, после которого её открывают заново — как обосновать решение.
- [ ] Посчитано, сколько расширений придётся проверять при обновлении.
- [ ] Написано, как забрать данные и сохранить адреса, если способ создания придётся сменить — стоимость ухода.

После запуска

- [ ] Кто-то следит, что ссылки и сервисы открываются.
- [ ] Открытая регистрация выключена, если внешние люди не публикуют, или включена модерация — доступ и роли. Обязательно для порталов, обновлённых со старых версий CKAN.
- [ ] Есть план закрытия: куда денутся данные и какие адреса будут перенаправлять — жизненный цикл.
- [ ] Каталог можно найти извне: регистрация в профильных реестрах и разметка, а не только новость на сайте.

Источники

- Карта возможностей
- Разобранные примеры
- Передовые каталоги

Разобранные примеры

Четыре карточки ниже доводят задачу до названной платформы. С какой строки начать — в главе «Типовые сценарии». Каждая записана семью разделами из главы «Как обосновать решение»: контекст, варианты, критерии, решение, почему, последствия, когда пересмотреть. Это образцы записи. Для другого города или службы разделы заполняют заново: чужая карточка тащит за собой чужое обязательство и чужую плату.

Числа жителей и наборов в примерах условные. Факты о платформах взяты из карточек и сверки октября 2026 года.

Коротко:

- В карточке одна развилка. Соседний отказ, который уже просили, назван отдельно, чтобы его не внедрили молча.
- В разделе «почему» каждый отвергнутый вариант проигрывает по названному критерию.
- Плата стоит рядом с выбором.
- «Когда пересмотреть» — событие, которое организация заметит. Новое решение — новая карточка.

Пример	Развилка	К чему приходят
Муниципалитет	свой портал, чужой каталог или статический сайт	публиковать на региональном портале
Регион	какая платформа соберёт города и отдаст страну	CKAN, ckanext-dcat и ckanext-harvest
Университет	репозиторий данных или портал открытых данных	Dataverse рядом с уже стоящим DSpace
Статистическая служба	один сайт или два каталога	показатели и микроданные на разных платформах

Как из задач получается список функций и короткий список платформ — в главах о функциях и о платформе.

Муниципалитет: не заводить свой портал

Контекст. Город на 40 тысяч жителей. Около тридцати наборов: бюджет, маршруты, адреса школ. Обновляют их два раза в год. Информатик администрации один, порталы он не сопровождал. Регион уже принимает наборы муниципалитетов и забирает описания в профиле DCAT-AP. Карта города лежит на региональном геопортале. Свой геопортал никто не просил.

Варианты. Свой CKAN. Облачный портал, Huwise или Socrata. Публикация наборов организацией на региональном портале, в том числе в типовом разделе, который регион уже ведёт для городов. Статический каталог на сайте города: JKAN или страницы с файлом описаний.

Критерии. Регион забирает описания без переписывания карточек, и файл совпадает с его профилем. Один человек без опыта порталов обновляет наборы два раза в год и через три года. Плата соразмерна тридцати наборам. Если критерии спорят, важнее первый: наборы должны дойти до региона.

Решение. Публиковать наборы организацией на региональном портале. Если регион не заводит чужие организации, город выкладывает статический каталог с файлом описаний в том профиле, который регион собирает.

Почему. Свой CKAN проигрывает по второму критерию. Выгрузка DCAT у него не в ядре, а в расширении ckanext-dcat, и обновление — это ядро плюс расширения. При одном человеке без опыта портал останется без обновлений через год-два. Облачный сервис проигрывает по третьему критерию. Страницу набора, лицензию, файл и выгрузку для региона он закроет, но эти функции уже есть на региональном портале. Подписка и будущий перенос адресов добавляют плату там, где своей установки нет. Чужой договор в карточку не переносят: сумму берут из предложения на эти тридцать наборов, как разобрано в главе о способе создания. Статический каталог остаётся запасным ходом, если регион не заводит чужие организации. У JKAN файл `data.json` объявляет схему DCAT-US 1.1, шаблон заполняет короткий список полей. Регион из этой карточки ждёт DCAT-AP, поэтому шаблон переделывают под его профиль до публикации. opendata.scot показывает другой случай: статический каталог собрали, когда государственного портала ещё не было. У города из этой карточки верхний портал уже есть.

Последствия. Город зависит от полей и расписания регионального портала. Отдельного домена с витриной не будет. Лицензию и постоянный адрес набор получает там же: регион их за город не выдумает. Карты остаются на геопортале. Если на набор нужно сослаться с городского сайта, страница ведёт на карточку региона, а не заводит вторую копию.

Когда пересмотреть. Наборов стало больше двухсот, появилась ставка на сопровождение портала, или регион перестал принимать муниципальные данные. Тогда открывают новую карточку. По этой записи CKAN молча не ставят.

Регион: собрать муниципалитеты и отдать страну

Контекст. Регион публикует свои открытые данные и обязан отдавать описания национальному порталу. Профиль — DCAT-AP плюс несколько своих обязательных полей. Одновременно регион собирает наборы муниципалитетов. Подрядчик есть, своей команды разработки нет. Часть городов уже ведёт каталог, часть будет публиковать прямо сюда, как в примере муниципалитета.

Варианты. CKAN с ckanext-dcat и ckanext-harvest. Piveau. Huwise. Socrata. Свой портал. Не собирать города и оставить их сами по себе.

Критерии. Файл выгрузки проходит валидатор профиля, который ждёт национальный портал, включая свои поля, и при смене профиля его чинит не единственный поставщик. Город без каталога публикует набор с файлом на региональном портале, город с каталогом отдаёт файл описаний, регион забирает его по расписанию. Подрядчик обновляет платформу на стенде. Если критерии спорят, совпадение с профилем страны важнее витрины и скорости запуска.

Решение. CKAN: ядро, ckanext-dcat и ckanext-harvest. Город без своего каталога публикует наборы организацией на региональном портале. Город со своим каталогом отдаёт файл описаний, регион его забирает. Свои поля национального профиля задаёт ckanext-scheming, в выгрузку их кладёт профиль ckanext-dcat. На scheming в тот же день пишут отдельную карточку расширения: что ставят, кто сопровождает, что делать, если расширение замолчит. Если страна принимает общий DCAT-AP без своих полей, scheming не ставят.

Почему. Socrata проигрывает по первому критерию, до разговора о цене. Сбор у неё есть: Catalog Connector забирает `data.json` и публичный ArcGIS Server. Штатный файл выгрузки — DCAT-US, а страна ждёт DCAT-AP.

Huwise проигрывает по тому же критерию. Сбор у неё тоже в самом сервисе: сборщики CKAN, Socrata, ArcGIS, `data.json` и других порталов Huwise, сбор из CSW включают по запросу в поддержку. Свои поля задаются шаблонами и попадают в выгрузку DCAT, только если поставщик предусмотрел соответствие. Подрядчик видит национальный профиль на пилоте у поставщика и сам его в репозитории не правит.

Piveau проигрывает по первому критерию в части своих полей. Собирать DCAT и публиковать наборы он умеет: на нём работает data.europa.eu, а форма ввода на октябрь 2026 года документирует спецификации DCAT-AP и DCAT-AP.de. Профиля этой страны в списке нет. Его добавляют отдельно, и дальше его сопровождает подрядчик. У CKAN общий профиль DCAT-AP 3 уже есть в ckanext-dcat, свои поля задаёт ckanext-scheming, в выгрузку их кладёт профиль того же расширения. Автоматические тесты идут на CKAN 2.10, 2.11 и 2.12.

Свой портал проигрывает по первому критерию тоже. Среди самописных порталов открытых данных выгрузку DCAT реестр нашёл у 24 из 1144. Сайт откроют люди. Национальный портал описания не заберёт.

Ход «оставить города сами по себе» проигрывает по второму критерию: страна муниципальные наборы не получит.

Сбор и выгрузка в этой сборке — два расширения, и они делают разную работу. Файл DCAT наружу и разбор чужого DCAT даёт ckanext-dcat. ckanext-harvest даёт очередь, расписание и экран ошибок. Последний выпуск harvest, 1.6.2, старше CKAN 2.12, но тесты основной ветки включают 2.12. В карточке записывают адрес репозитория, который ставят, и проверку сбора на стенде той версии ядра, которую выбрали. Подробности — в главе про CKAN.

Последствия. Подрядчик при каждом обновлении CKAN проверяет ckanext-dcat, ckanext-harvest и, если он стоит, ckanext-scheming. Ошибки сбора смотрит человек. В карточке копии стоит ссылка на страницу набора у города, иначе один набор считается и у города, и у региона, и у страны. В ядре CKAN ни выгрузки, ни сбора нет: это штатная сборка. Та же связка уже стоит на живом портале: у data.gov.gr на CKAN 2.11.3 в сверке 13 октября 2026 года среди плагинов были scheming, dcat и harvest.

Когда пересмотреть. Национальный портал сменит профиль, или ckanext-dcat либо ckanext-harvest перестанут проверяться на новой версии CKAN. Тогда ищут замену расширения или остаются на последней версии ядра, которая ещё получает исправления безопасности. Как выбрать из этих ходов — в главе об обосновании. Если страна обяжет регионы работать на одной государственной платформе, свою сборку закрывают по плану переноса адресов.

Университет: репозиторий данных, а не «ещё один CKAN»

Контекст. Университет уже ведёт репозиторий публикаций на DSpace. Гранты требуют выкладывать данные к статьям с DOI, версией и строкой цитирования, которую можно выгрузить в BibTeX. Публикуют научные группы, библиотека готова модерировать. Отдельной группы под новое приложение нет: систему ведёт та же библиотека. Открытых данных администрации немного, им место на странице университета или на портале органа власти. Федерация DCAT-AP государственного портала на первом шаге не обязательна.

Варианты. Доработать стоящий DSpace. Поставить рядом Dataverse. Поставить рядом InvenioRDM. Не ставить ничего и отправлять авторов в Zenodo. Поставить CKAN и добавить расширение ckanext-doi.

Критерии. DOI, версии, строка цитирования и выгрузка BibTeX есть в ядре той версии, которую поставят. Библиотека модерирует запись, файл можно закрыть при открытом описании. У университета свой префикс DOI. Уже стоящий репозиторий публикаций этой карточкой не заменяют, и сопровождает новую систему библиотека, без отдельной команды. Если критерии спорят, цитирование в ядре важнее желания обойтись одной программой.

Решение. Dataverse рядом с DSpace. Общая точка входа — на сайте библиотеки. В описании набора данных стоит DOI статьи, в описании статьи — DOI набора. Группам с единичными наборами по-прежнему можно пользоваться Zenodo.

Почему. CKAN проигрывает по первому критерию. DOI у него — расширение ckanext-doi, на CKAN 2.11 и 2.12 оно не тестировалось. Версий набора в ядре нет. OAI-PMH, которым научные агрегаторы забирают описания, в ядре тоже нет.

DSpace проигрывает по тому же критерию, хотя DOI в ядре включается настройкой. До версии 10.1 включительно готовой ссылки и выгрузки в BibTeX и RIS нет: в ядре только метатеги для Google Scholar, кнопку «Cite» переносят из DSpace-CRIS. В DSpace 10 появился тип записи «набор данных», но в репозиториях на DSpace наборы остаются малой долей записей, а фильтр по типу в поиске по умолчанию выключен. Дорабатывать публикационный репозиторий до репозитория данных библиотека сможет, когда выгрузка цитирования появится в ядре. До этого дня данные идут в отдельную систему.

Zenodo проигрывает по третьему критерию. DOI, версии и цитирование там есть, установка не нужна, на одну запись можно положить до 50 ГБ и 100 файлов. Префикс DOI общий, модерацию групп библиотека не ведёт.

InvenioRDM закрывает DOI, версии и BibTeX в ядре. Карточка платформы отдаёт его организации, которая готова сопровождать Python-приложение. У этой библиотеки такой готовности нет, поэтому в коротком списке он не остаётся. Университету с отдельной командой и с обязанностью самому отдавать DCAT он подходит: выгрузка DCAT у InvenioRDM в ядре, у Dataverse её нет. В этой карточке федерация государственного портала не обязательна, и выгрузка DCAT решение не перевешивает.

Последствия. Две системы у библиотеки. Связь «статья — данные» держат идентификаторами в описаниях. Общего поиска по ним эта карточка не обещает. Нужен договор с DataCite, если его ещё нет. Самостоятельный член платит 2000 евро в год и инфраструктурный сбор от 200 до 700 евро, участник консорциума — только сбор.

Университеты обычно получают префикс через библиотечный консорциум. Пока префикса нет, группы пользуются Zenodo. Боевые DOI с тестовой учётной записи не выпускают: ошибку в рабочей системе трудно исправить, DOI не удаляются.

Когда пересмотреть. Сопровождать две системы некому — свой Dataverse закрывают и договариваются о Zenodo или общем репозитории консорциума. Вышла версия DSpace, в которой строка цитирования и BibTeX есть в ядре наравне с Dataverse, — открывают новую карточку о слиянии. Появилась обязанность самому отдавать описания в DCAT-AP — карточку открывают заново и снова смотрят InvenioRDM.

Статистическая служба: показатели и микроданные — разные каталоги

Контекст. Национальная статистическая служба ведёт две полки, которые часто пытаются закрыть одним сайтом. Первая — показатели с разрезами по времени и территории и обмен с международными организациями по SDMX. Вторая — обезличенные файлы обследований: описание открыто, файл по заявке, описание в духе DDI.

Варианты. Один самописный сайт «вся статистика». Два каталога: платформа показателей и NADA для микроданных. Только платформа показателей, а микроданные рассылать почтой.

Критерии. Программа получает срез показателя по году и территории, международный партнёр забирает SDMX. У обследования открытое описание, уровни доступа и описание в DDI. Служба не пишет оба контура сама. Проигрыш любого из первых двух критериев значит, что каталог к работе не готов.

Решение. Два каталога. Микроданные — на NADA. Платформу показателей выбирает следующая карточка. На сайте службы обе точки входа стоят рядом, и подписи прямо говорят, чем полка показателя отличается от полки обследования. Показатели внутри NADA не включают. С выпуска 5.6 платформа умеет их описывать, но возможность ещё не делает каталог гибридом: в библиотеке микроданных Всемирного банка все 7 145 записей — обследования, а показатели банк публикует на другой платформе.

Почему. Самописный сайт проигрывает по третьему критерию и обычно по первому. Среди каталогов показателей самописных 44%, это 1505 витрин по выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года). Программный доступ есть у 37% каталогов этого типа. У самописных витрин API отмечен у 8%. SDMX и уровни доступа в таком сайте придётся написать и не бросить. Почта вместо каталога проигрывает по второму критерию: исследование нельзя найти и на него нельзя сослаться, пока его нет в каталоге. NADA закрывает уровни доступа и DDI в ядре. Dataverse ближе университетскому архиву: DOI и OAI-PMH у него в ядре, а семи типов доступа нет.

Последствия. Два сопровождения и два адреса. Пользователь путается, если подписи на сайте не различают полки. Живой образец такой раскладки — мексиканский INEGI: показатели, микроданные на NADA, геопортал и открытые данные стоят рядом и записаны разными каталогами. NADA ставят на домен службы и закладывают сопровождение в бюджет, не только установку. Из 59 неработающих каталогов микроданных 47 — брошенные установки NADA после проектов поддержки. В NADA нет OAI-PMH. DOI она не регистрирует: поле для уже выпущенного DOI есть, выпуск — отдельная задача. Если службе нужен обмен с

архивами вроде CESSDA или свой префикс DOI, это следующая карточка, а не повод переехать на Dataverse молча. Nesstar для нового каталога не рассматривают: с версии 4.0 в 2015 году его объявили устаревшим, к 2022 году сняли.

Когда пересмотреть. NADA перестанут выпускать — преемника ищут заранее, не ждут, пока сайт останется без обновлений. На октябрь 2026 года этого события нет: выпуск 5.7.1 вышел 4 октября. Обмен с архивами социальных данных станет обязанностью — открывают карточку про OAI-PMH и DOI.

Какая платформа показателей

Контекст. Предыдущая карточка уже развела показатели и микроданные. Эта выбирает только полку показателей. Обмен по SDMX обязателен. Своей большой команды разработки нет, бюджет на консультантов ограничен.

Варианты. Только PxWeb. Только .Stat Suite. PxWeb для витрины и отдельно .Stat Suite или SDMX-RI как точка обмена.

Критерии. Есть API и выборка по разрезам. Международный партнёр получает SDMX со структурами и справочниками. Служба тянет сопровождение того, что поставила. Если знакомый интерфейс обмен не закрывает, обмен важнее интерфейса.

Решение. Если консультанты по SDMX есть и обмен — главная обязанность, показатели ставят на .Stat Suite: служба SDMX у неё в ядре, внутри используется SDMX-RI. Если команда уже умеет PxWeb, а обмен нужен рядом, витрину ставят на PxWeb. Новые установки — сразу на PxWebApi 2. Точку SDMX добавляют отдельным продуктом, .Stat Suite или SDMX-RI. Два статистических продукта здесь обычная схема: витрина для людей и служба для обмена.

Почему. Один PxWeb проигрывает по второму критерию. Полноценной службы SDMX у него нет: старый API отдаёт файл SDMX-ML 2.0, в PxWebApi 2 формата SDMX нет. Разбор — в главе о каталогах показателей. Одна .Stat Suite закрывает API и SDMX и проигрывает по третьему критерию, если консультантов нет: без них её ставят редко. Самописную витрину поверх этой пары не делают. Порталы ЕЦБ и Статистики Канады стоят поверх API и массовой выгрузки. У службы из этой карточки такой команды нет.

Последствия. При выборе «PxWeb плюс точка обмена» сопровождают два статистических продукта, не считая NADA. Переход на PxWebApi 2 меняет адреса запросов. Старые ссылки переводит конвертер, старый API держат, пока пользователи не переехали, и заранее называют срок отключения. У .Stat Suite обозреватель данных и служба SDMX — разные модули, обновляют оба.

Когда пересмотреть. Международный обмен уйдёт с SDMX на другой обязательный протокол. Команда, которая умела только PxWeb, получит консультантов по SDMX и решит снять отдельную точку обмена — это новая карточка.

Что общего у примеров

Отказ записан так же явно, как выбор. Плата названа. Пересмотр привязан к событию.

Пример	Что отвергли и по какому критерию	Чем платят	Когда открыть карточку заново
Город	свой CKAN — некому обновлять; облако — те же функции уже есть у региона	зависимость от полей и расписания региона; витрина на своём домене не появляется	больше двухсот наборов, ставка на сопровождение, регион перестал принимать данные
Регион	Socrata — файл DCAT-US; Huwise и Piveau — свой профиль страны сопровождает поставщик или подрядчик; свой сайт — выгрузка DCAT редка	расширения, которые проверяют при обновлении, и человек на ошибки сбора	смена профиля, расширение не проверяется на новой версии ядра, страна обязала сидеть на одной платформе
Университет	CKAN и один только DSpace — нет цитирования в ядре; Zenodo — нет своего префикса	две системы и договор с DataCite	некому вести две системы; у DSpace появилась выгрузка цитирования; понадобился свой DCAT-AP
Статслужба	один самописный сайт — API и SDMX придётся писать самим; почта — обследования не найти	два адреса у показателей и микроданных, а при PxWeb — ещё точка обмена SDMX	NADA перестанут выпускать; обмен уйдёт с SDMX

Чужая строка этой таблицы другому владельцу не подходит. Региону нужна сборка, которую городу сопровождать некому. Университетский Dataverse не становится порталом открытых данных администрации. NADA с выпуска 5.6 не заменяет каталог показателей.

Источники

- Как обосновать решение, принципы выбора функций, от функций к платформе, свои силы, свободное ПО или облачный сервис — шаблон карточки, матрица региона и почему чужой договор в карточку не переносят.
- CKAN: ядро и расширения — ckanext-dcat, ckanext-harvest 1.6.2 и тесты на CKAN 2.12, ckanext-scheming, ckanext-doi; сверка расширений 11 октября 2026 года. Живой список плагинов data.gov.gr — 13 октября 2026 года, в главе о карточках платформ.
- Экспорт метаданных — файл `data.json` у JKAN объявляет схему DCAT-US 1.1 и заполняет короткий список полей; штатный файл Socrata — DCAT-US; у ckanext-dcat по умолчанию профиль DCAT-AP 3.
- Сборщики Huwise, проверено 14 октября 2026 года; Catalog Connector у Socrata.
- Форма ввода Piveau — заявленная поддержка DCAT-AP и DCAT-AP.de, проверено 14 октября 2026 года. Сбор как основная работа ядра — в главе о федерации.
- Научные репозитории — DOI, версии и цитирование Dataverse, InvenioRDM и DSpace 10.1; лимиты Zenodo. Взносы DataCite — в главе об идентификаторах.
- Каталоги показателей — 1505 самописных витрин, API у 37% каталогов типа и у 8% самописных, PxWeb и SDMX. Каталоги микроданных — NADA 5.7.1 от 4 октября 2026 года, 47 из 59 неработающих каталогов, снятие Nesstar. Сравнение типов и гибриды — INEGI и 7 145 обследований библиотеки Всемирного банка.
- opendata.scot — статический каталог, пока не было государственного портала; разбор — в главе о публикации без портала.
- Доли DCAT у самописных порталов и доли API у каталогов показателей — выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года.

ЧАСТЬ VII.

Как публикуют данные

Процесс публикации, стандарты метаданных, федерации, лицензии и находимость.

1. Процесс публикации
2. Стандарты метаданных
3. Федерации и агрегация
4. Лицензии и условия доступа
5. Находимость

Процесс публикации

Набор опубликован, когда посторонний человек находит карточку, открывает файл или видит, как его запросить, читает лицензию и может назвать, кто ответит за строки. Файл, выложенный один раз к отчёту, этого не делает: карточка есть, а обновлять её некому.

Порядок один и тот же для портала, статического сайта и репозитория. Меняется инструмент на шаге «публикация». Жизнь самого каталога — запуск, смена платформы, закрытие — разобрана в главе о жизненном цикле.

Коротко:

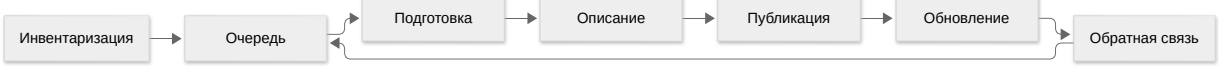
- Сначала записывают, что у организации уже есть. Потом решают, что из этого публиковать и в каком порядке.
- Публикация — три проверки: карточка в поиске, файл или сервис открывается, описание есть в общей выгрузке.
- У набора назван человек, который ответит на ошибку в строках.
- Набор, на который уже могли сослаться, не снимают молча: адрес остаётся и объясняет, что случилось.
- Просьба о недостающем наборе возвращается в очередь.

Когда набор можно считать опубликованным

Что видно снаружи	Что должно быть правдой
Карточка находится в поиске	в поиск попал выпущенный набор, а не черновик
У файла есть дата	дата относится к строкам и не старше обещанной частоты
Где-то написано «свободное использование»	лицензия стоит на наборе ссылкой и попадает в выгрузку
В подвале сайта есть `info@`	у набора названы имя и адрес того, кто ответит по существу
Набор исчез из каталога	прежний адрес открывается и объясняет, почему выпуска больше нет или чем его заменили

Адрес, лицензия, открытый файл или честный отказ, дата и общая выгрузка уже стоят в чек-листе требований. Ниже — порядок, которым к этому списку приходят.

Семь шагов



Владелец данных отвечает за строки. Публикатор выкладывает файл и карточку. Редактор ведёт очередь и выгрузку. Полное разделение ролей и то, что ломается при пустой роли, — в таблице ниже.

Шаг	Готово, когда	Кто закрывает шаг
Инвентаризация	есть список того, что организация уже производит, включая то, что публиковать нельзя	владельцы данных вместе с редактором каталога
Очередь	у каждого пункта записано, почему он раньше остальных, и список переживает год	редактор каталога
Подготовка	файл открывается без платной программы, колонки названы, люди обезличены, лицензия согласована	владелец данных и юрист
Описание	заполнен минимум карточки и поля, без которых набор вашей дисциплины нельзя понять	публикатор по образцу владельца
Публикация	набор в поиске, файл или сервис открывается, строка есть в общей выгрузке	публикатор; при проверке куратором — второй человек
Обновление	частота записана, просрочка видна, адрес при замене файла не меняется	владелец данных
Обратная связь	письмо об ошибке доходит до владельца, а просьба о новом наборе попадает в очередь	владелец данных и редактор

Инвентаризация

Инвентаризация — список того, что уже существует и у кого лежит: реестры, отчёты, слои, обследования, выгрузки учётной системы. Это список активов, а не план «что хочется выложить» и не сам каталог. Пустой список на этом шаге полезнее портала без данных: публиковать пока нечего.

Колонок хватает нескольких: название, где лежит сейчас, кто отвечает за строки, есть ли в данных люди, как часто строки меняются, требует ли закон или грант публикацию, просил ли кто-то этот набор. Для органа власти заготовка часто уже есть в учётной системе. Для университета — в плане управления данными гранта.

Федеральное ведомство США по меморандуму OMB M-25-05 от 15 января 2025 года ведёт такой список по всем активам данных, кроме систем национальной безопасности. Новый актив вносят не позднее чем через 90 дней после того, как его создали или нашли. Весь перечень сверяют по ходу работы и не реже раза в год. Файл публикуют в JSON по адресу `www.[agency].gov/data.json`, в схеме DCAT-US 3.0. Сам перечень скачивается целиком. Файл данных есть не у каждой записи: решение отдать данные принимают отдельно, а закрытый актив в перечне остаётся, с пометкой, как к нему проситься. Срок привести перечень к схеме 3.0 и включить в него все активы истёк 30 сентября 2026 года; как выглядит такой перечень через две недели после срока, видно на примере ниже.

14 октября 2026 года файл Администрации общих служб (GSA) по адресу `www.gsa.gov/data.json` открывался и перенаправлял на `open.gsa.gov/data.json`. В нём 344 записи, схема объявлена как DCAT-US 3.0. У всех 344 заполнено контактное лицо: имя и почта. Открытых записей 310, закрытых от публикации — 33, с ограниченным доступом — одна. У закрытой записи «Concur - Reporting Authorization Model» нет файла, а в условиях доступа сказано, что к публичному выпуску набор не предназначен и нужно писать издателю. Так перечень отличается от витрины скачиваний: видно, что актив есть, и видно отказ.

Ссылка на лицензию при этом есть у 320 из 344 записей. У 24 открытых записей поле пустое. Перечень уже отделяет закрытое от открытого, а готовой публикацией он от этого не становится.

Срок сам по себе файл не переписывает. У Избирательной комиссии США `eac.gov/data.json` 14 октября 2026 года по-прежнему объявлен схемой DCAT-US 1.1 и содержит 11 записей. Как национальный каталог забирает такие файлы — в главе «Когда портал не нужен».

Очередь

Сначала то, что требуют закон, грант или вышестоящий портал, и то, о чём уже просят. Остальное может подождать. Очередь записывают, иначе каждый год начинают с нуля.

В ЕС вперёд очереди стоят наборы высокой ценности: шесть категорий, которые с 9 июня 2024 года отдают бесплатно, через API и массовой выгрузкой. В США план открытых данных ведомства должен описать, как житель просит поднять конкретный актив в очереди. Меморандум среди поводов называет уже идущий выпуск, прямую пользу для людей, частые запросы по закону о доступе к информации и данные, на которых проверяют расчёты и модели. Справочник, который никто не спрашивал, в этой очереди последний.

Подготовка

Формат, который открывается без платной программы. Словарь полей. Обезличивание, если в данных есть люди. Лицензия, с которой согласен юрист. Как выбрать формат — в главе о форматах, как выбрать лицензию — в главе о лицензиях. Форма и скрипт этот шаг не заменяют: они выкладывают то, что человек уже решил.

Описание

Минимум из главы о полноте метаданных: название, кто опубликовал, лицензия, дата, постоянный адрес. Плюс поля, без которых набор вашей дисциплины нельзя понять: период и территория, переменные обследования, система координат слоя. В российском паспорте набора по методическим рекомендациям 4.0 ответственное лицо и его контакты входят в сам бланк. Как устроен паспорт — в главе о стандартах метаданных. Международный сборщик этот бланк не читает. Человеку, который ищет, кому написать, поле всё равно нужно.

Публикация

Набор виден в поиске, файл или сервис открывается, описание отдаётся в общей выгрузке. Черновик в поиск не попадает. Кто имеет право нажать «опубликовать», задают роли. В научном репозитории этот шаг часто делает второй человек: автор сдаёт набор, куратор выпускает. Как устроены черновик, роли и проверка — в главе о доступе.

Выложить карточку можно тремя способами.

Способ	Когда его хватает
Форма на портале	наборов десятки, правят их два человека, в учётной системе этих полей нет
Скрипт из системы, где данные уже лежат	наборов сотни, карточка собирается из тех же полей, которые человек заполнил бы руками
Файл описаний, который забирает чужой каталог	своего портала нет или национальный каталог сам ходит за файлом

Руководство data.gouv.fr советует программный интерфейс, когда набор обновляют регулярно, и форму, когда выкладывают редко. Скрипт не назначает лицензию и признак «можно публиковать»: эти поля в системе заполняет человек до запуска. У платформы udata, на которой работает data.gouv.fr, есть отдельная ловушка. Набор, созданный через API, станет публичным, если при каждом запросе не передать флаг закрытости. Черновик в скрипте задают явно.

Обновление

У набора указано, как часто меняются строки. Рекомендации W3C «данные на вебе» прямо связывают это с процессом: публикация в вебе — часть выпуска новой версии, и за выпуск отвечает конкретный человек. Проверка короткая. Частота названа, и последний файл не старше даты, которую эта частота обещает. Каталог, который сам считает срок и пишет владельцу набора, — Humanitarian Data Exchange. Как устроен счёт — в главе о хороших практиках.

В перечне GSA частота «раз в квартал» стоит у 121 записи. В файле это повторяющийся интервал `R/P3M`. У 112 из них дата в поле `modified` — 2020 год или раньше. Ещё у 89 записей частота не заполнена. Двадцать ссылок на файлы ведут на имя `www.asap.gsa.gov`, которое 14 октября 2026 года не открывалось. Среди них набор «Networx Business Volume FY2013, 3rd Qtr»: квартальная частота и дата изменения 12 июня 2019 года. Сам перечень при этом правят. Его собственная запись изменена 5 октября 2026 года. Обновление карточек отстало от обновления файла перечня.

Набор, на который уже могли сослаться, не удаляют молча. Руководство GSA по DCAT-US 3.0 говорит то же: снятие без объяснения выглядит как сокрытие. Срок хранения пишут заранее, в описании или в поле прав. У записи остаётся контакт, чтобы можно было спросить, почему набор сняли. Если устарел формат, до полного снятия пробуют выложить простой статичный файл. Три исхода, и у каждого адрес набора сохраняется.

- Файл заменяют, дата строк меняется.
- Выходит новая версия, старая остаётся находиться.
- Выпуск закрыт. На прежнем адресе написано, почему и чем пользоваться вместо него.

Молчаливое удаление разобрано среди антипаттернов. Страница, которая остаётся на месте снятого набора с DOI, — в главе об идентификаторах.

Пометка «закрыто» без проверки ссылки тоже дырявая. У 14 из 33 закрытых записей GSA в описании всё же указана ссылка на скачивание, хотя текст доступа говорит, что набор не для публичного выпуска.

Обратная связь

Нужно место, куда написать об ошибке, и место, где попросить недостающий набор. Для портала органа власти это часть задачи. Письмо доходит до владельца данных.

Рекомендации W3C делят это на два шага. Способ написать легко найти на странице набора. Ответы видны следующим читателям, чтобы одну и ту же ошибку не присылали заново. На data.gouv.fr это обсуждение у карточки. Как портал отдаёт обсуждения в общей выгрузке — в главе о хороших практиках.

Просьба «опубликуйте вот это» без очереди остаётся в почте. Её записывают следующим пунктом очереди и отвечают, взяли её в работу или нет.

Кто за что отвечает

Роль	За что отвечает	Чего не делает	Если роли нет
Владелец данных	содержимое, срок обновления, ответ на ошибку по существу	не выбирает платформу и не верстает карточку наугад	набор выложили один раз, на письмо некому ответить
Публикатор	выкладывает файл и заполняет карточку по образцу	не назначает лицензию в день выкладки	владелец правит портал сам, и карточки расходятся
Редактор каталога	карточки устроены одинаково, очередь не застревает, выгрузка жива	не подменяет владельца данных	каждый год инвентаризацию начинают с нуля
Юрист и специалист по персональным данным	лицензия и обезличивание до публикации	не появляются впервые после жалобы	лицензию пишет публикатор в день выкладки
Администратор платформы	сервер, обновления программы, роли, копии	не решает, какой набор публиковать	каталог открывается, пока не кончилась версия или место на диске

В маленькой группе один человек может закрывать две соседние роли. Все пять он не закрывает: тот, кто отвечает за строки, не должен быть единственным, кто умеет поднять сервер. Так разделены координаторы открытых данных Нью-Йорка и команда самого портала. Как это записано — в главе о жизненном цикле.

Что меняет тип каталога

Семь шагов остаются. Тип двигает один или два из них.

Тип	Какой шаг другой	Что меняется
Портал открытых данных	инвентаризация и очередь	список того, что ведомство уже ведёт; впереди закон и просьбы
Научный репозиторий	публикация и обновление	набор выпускает второй человек; правка файла — новая версия
Геопортал	подготовка	система координат и описание ISO; в общий каталог кладут ссылку на это описание
Каталог показателей	обновление	ряд выходит по календарю, методика лежит рядом с рядом
Каталог микроданных	подготовка и публикация	уровень доступа решают до выкладки; описание открыто, даже когда файл закрыт, и написано, как его запросить
ML-каталог	подготовка и описание	деление на части и происхождение фиксируют до публикации
Публикация без портала	публикация	те же шаги, на последнем — страница и файл

Руководство GSA для геоданных просит не переписывать подробное описание ISO внутрь DCAT. В записи каталога ставят ссылку на уже готовый файл метаданных и на словарь полей. Порог, с которого нужен свой портал, — в главе «Когда портал не нужен».

Пример: наборы уже лежат в системе

Контекст. У регионального ведомства около двухсот наборов. Они уже есть в учётной системе: название, дата, подразделение. Национальный портал забирает описания в DCAT. Форму на своём портале заполняют сотрудники, у которых нет времени на вторую карточку. Каталог отстаёт от системы примерно на квартал.

Варианты. И дальше заполнять форму руками. Скрипт, который создаёт и обновляет карточку из учётной системы. Файл описаний на сайте ведомства, который забирает сборщик, как американский `data.json`.

Критерии. Каталог не отстаёт от системы на квартал. Лицензию и признак «можно публиковать» скрипт не назначает сам. На ошибку в строках отвечает человек.

Решение. Карточки создаёт и обновляет скрипт из полей системы. Лицензию, признак публикации и имя владельца в системе заполняет человек до запуска. Форма остаётся для наборов, которых в системе нет.

Почему. Ручная форма на двести наборов отстаёт в первый же квартал. Файл для сборщика закрывает задачу, когда своего портала нет или национальный каталог ходит за файлом на сайт ведомства. Здесь портал с ролями подразделений уже стоит, поэтому скрипт пишет в него напрямую: сборщик эти роли не ведёт. Скрипт без полей лицензии и владельца выложит пустые карточки быстрее, чем люди. Критерий «отвечает человек» такой скрипт проигрывает.

Последствия. В учётной системе появляются три поля, которых там могло не быть: лицензия, признак публикации, владелец. Сломанный скрипт молчит, поэтому у него тот же контроль срока, что у набора. Просрочка видна редактору.

Когда пересмотреть. Национальный портал сменит профиль, и скрипт начнёт отдавать файл, который тот не примет. Учётную систему заменят. Большая часть новых наборов перестанет проходить через систему, и форма снова станет основным путём.

Как записать свою карточку — в главе «Как обосновать решение».

Один набор: расписание автобусов

Инвентаризация находит расписание в транспортной системе. В очереди оно выше альбома фотографий остановок: его уже просят, и его может ждать региональный портал. Подготовка: CSV с колонками из словаря, без объединённых ячеек, лицензия, с которой согласен юрист. Описание: название, перевозчик как издатель, период действия, лицензия ссылкой, дата строк, имя того, кто ответит за ошибку в маршруте. Публикация: карточка в поиске, файл открывается, строка есть в общей выгрузке. Обновление: при смене маршрута меняют файл и дату, адрес карточки остаётся. Обратная связь: сообщение «ошибка в расписании» уходит перевозчику. Та же рекомендация W3C разбирает остановки на похожем примере: частота записана и словами, и в поле, которое читает программа.

Если наборов таких двадцать и правят их два человека, хватит таблицы и статического сайта. Если наборов двести и правят двенадцать ведомств, нужна платформа с ролями. Порог — в главе «Когда портал не нужен».

Как понять, что процесс живой

Это можно проверить снаружи, не собирая редакцию.

- У взятого наугад набора в карточке есть имя или команда и адрес, а не только ящик всего портала.
- Дата файла не старше обещанной частоты. Пустая частота значит, что для этого набора обновление не задано.
- Черновик поиском не находится.
- Адрес набора, который снимали, открывается и объясняет, что с ним стало.
- Есть куда попросить недостающий набор. У портала органа власти эта просьба видна как очередь или как публичное обсуждение.
- Сам перечень существует как документ — таблица или файл вроде `data.json`, — а не только как память редактора.

Тот же список, короче, — в чек-листах.

Источники

- Data on the Web Best Practices, W3C, 31 января 2017 года — практика 21 (данные актуальны, частота названа, за выпуск отвечает человек), практика 22 (если данных нет, написано, как их получить), практики 29 и 30 (отзыв можно оставить, и он виден следующим читателям); пример с остановками и полем частоты
- OMB, меморандум M-25-05, 15 января 2025 года — инвентаризация всех активов, кроме систем национальной безопасности; новый актив не позднее 90 дней; сверка не реже раза в год; файл `data.json` в схеме DCAT-US 3.0; срок 30 сентября 2026 года; в плане открытых данных — способ просить поднять актив в очереди
- GSA, DCAT-US v3.0 Implementation Guide, редакция 1.1 от 9 сентября 2026 года — контакт обязателен; снятие набора сопровождают объяснением; для геоданных в DCAT ставят ссылку на готовое описание ISO, а не переписывают его
- GSA, How to get your Open Data on Data.gov — национальный каталог забирает файл описаний, файлы ведомства остаются у ведомства
- guides.data.gouv.fr: опубликовать набор — при регулярном обновлении карточка уходит через API, при редком хватает формы
- Проверка 14 октября 2026 года: `www.gsa.gov/data.json` перенаправляет на `open.gsa.gov/data.json`, 344 записи, схема DCAT-US 3.0, контакт у всех 344, уровни доступа 310 / 33 / 1, лицензия у 320, квартальная частота у 121 и у 112 из них `modified` не новее 2020 года, частота пустая у 89, 20 ссылок на `www.asap.gsa.gov` (имя не открывается), запись самого перечня изменена 5 октября 2026 года; `eac.gov/data.json` — 11 записей, схема DCAT-US 1.1
- Чек-лист требований, хорошие практики, полнота метаданных, лицензии, доступ и роли

Стандарты метаданных

Стандарт метаданных — общий язык описания. Каталог отдаёт карточки так, чтобы их понял другой портал, поисковик или программа. Свой набор полей понятен только тем, кто читал вашу инструкцию.

Главный вывод главы: берут язык, который уже читает адресат, и часто два языка сразу. Федерация читает профиль DCAT своей страны, поисковик — разметку schema.org на странице, регистрация DOI — схему DataCite. Эти языки описывают один и тот же набор и не подменяют друг друга. Черновик следующей версии и «JSON по мотивам DCAT» для обмена не годятся.

Как платформа отдаёт файл и какой профиль в ней включён по умолчанию — в главе об экспорте. Какие поля заполнять, чтобы набор можно было использовать, — в главе о полноте метаданных. Здесь — какой язык для какого каталога и как узнать его по файлу.

Описание набора и устройство данных

Стандарты вокруг данных делятся на два вида, и их часто путают.

Стандарт метаданных описывает набор снаружи: что это, кто издатель, за какой период, где скачать, на каких условиях. Его читает каталог, агрегатор, поисковик. DCAT, DataCite, ISO 19115, schema.org — стандарты метаданных.

Стандарт данных описывает, что внутри файла: какие колонки или поля, что они значат, как кодируются значения. Его читает программа, которая обрабатывает сами данные. OCDS для закупок, GTFS для расписаний, Darwin Core для наблюдений видов — стандарты данных. Они разобраны в главе о том, что публикуют.

Каталогу нужен первый вид всегда, когда описания забирает кто-то снаружи. Второй нужен, когда в отрасли он есть: набор закупок в OCDS сразу понимают программы анализа контрактов, а свой CSV про контракты каждый разбирает заново. Связь карточки со стандартом файла в DCAT записывают свойством `dct:conformsTo`. Словарь колонок — отдельная задача, она разобрана в разделе о словаре полей.

Некоторые стандарты делают обе работы. SDMX описывает и показатель, и структуру ряда. DDI описывает и обследование, и переменные. Croissant описывает и набор для машинного обучения, и его файлы с полями. STAC описывает и коллекцию снимков, и отдельные файлы.

Если нужного стандарта не знаете, поищите в каталогах стандартов: FAIRsharing для науки, Data Standards Directory для государственных данных, перечень data.gov для США.

Четыре слоя, которые смешивают

В разговоре «мы поддерживаем DCAT» смешивают четыре разные вещи. Пока они не разделены, нельзя понять, пройдёт ли выгрузка проверку.

Слой	На какой вопрос отвечает	Пример
Словарь	что означает поле	DCAT, schema.org, DataCite, ISO 19115
Профиль	какие поля обязательны и из какого списка берётся значение	DCAT-AP, DCAT-US, GeoDCAT-AP
Справочник	какое именно значение можно записать	темы data theme, лицензии, коды языков
Запись и протокол	в каком файле приходит ответ и как его спросить	JSON-LD, Turtle, XML; OAI-PMH, CSW

Словарь задаёт смысл. DCAT говорит, что у каталога есть наборы, у набора — способы получения, у способа получения — ссылка на файл. Профиль сужает словарь под страну или отрасль: оставляет нужные поля, помечает их обязательными или рекомендуемыми и привязывает к справочникам. DCAT-AP и DCAT-US — профили одного DCAT, а не два разных языка. Национальный профиль, например итальянский DCAT-AP_IT или швейцарский DCAT-AP-CH, садится поверх DCAT-AP и добавляет свои поля.

Запись — это файл. Одно и то же описание DCAT можно отдать текстом Turtle, XML или JSON-LD. Для человека разницы нет, для программы есть: она просит нужный вид заголовком `Accept` или берёт адрес с нужным окончанием. Протокол — способ спросить каталог: OAI-PMH, CSW, OGC API Records. Один CSW умеет ответить и карточкой ISO 19115, и описанием DCAT, если это умеет программа. Подробнее о способах отдачи — в главе об экспорте.

Отсюда практическое следствие. Портал может честно отдавать DCAT и при этом не проходить валидатор профиля, который ждёт сборщик. 13 октября 2026 года выгрузка data.gov.gr была в профиле DCAT-AP 3, а валидатор DCAT-AP 3.0.1 нашёл две ошибки: формат файла записан словом «CSV», а не ссылкой на справочник, и у контрольной суммы не указан алгоритм. Разбор — в главе об экспорте.

Что брать за основу

Каталог	Что отдавать наружу	На что опираться	Кого этим не закрыть
Портал открытых данных	профиль страны и schema.org на странице	DCAT-AP 3.0.1 в ЕС; DCAT-US 3.0 в США; в России — паспорт методических рекомендаций 4.0	поисковик не читает только DCAT, сборщик не читает только разметку страницы
Геопортал	ISO 19115; для портала открытых данных ещё GeoDCAT-AP	ISO 19115, в ЕС — INSPIRE и GeoDCAT-AP 3.1.0; файлы снимков — STAC	карточка слоя не заменяет описание файлов, STAC не заменяет инфраструктуру страны
Научный репозиторий	DataCite для DOI, Dublin Core в OAI-PMH, schema.org на странице	DataCite 4.7; для передачи папки с файлами — RO-Crate	DOI не регистрируют записью DCAT
Каталог показателей	SDMX для партнёров; StatDCAT-AP, если таблицы должны попасть на портал открытых данных	SDMX 3.1, на практике часто ещё 2.1; таблицы в браузере — JSON-stat	SDMX не делает набор находимым в Google Dataset Search
Каталог микроданных	DDI на обследование и schema.org на странице	DDI Codebook 2.5; новая 2.6 — с 15 апреля 2026 года	DDI не заменяет условия доступа и DOI
ML-каталог	карточка для человека и Croissant для программы	Croissant 1.1 от 29 января 2026 года; площадки в 2026 году ещё отдают 1.0	Croissant не описывает показатель и не заменяет лицензию
Реестр словарей	SKOS, для элементов данных — ISO/IEC 11179	SKOS; ISO/IEC 11179-3:2023	словарь терминов не заменяет каталог наборов

Два языка сразу — обычная работа, а не запасной вариант. Внутренняя форма каталога может быть богаче профиля: свои поля для редакции остаются у вас. Наружу уходит профиль, который назвал сборщик.

DCAT и профили

DCAT — словарь W3C для каталогов данных. Первая рекомендация вышла 16 января 2014 года, вторая — 4 февраля 2020 года, третья — 22 августа 2024 года. Версия 3 описывает каталог, набор, способ получения, сервис данных, серию наборов и запись каталога. В ней появились серии отдельно публикуемых наборов и свойства версии, включая контрольную сумму файла. Как эти части выглядят на живой карточке, разобрано в главе об устройстве каталога.

Сам DCAT почти ничего не объявляет обязательным. Обязательность появляется в профиле. Поэтому фраза «у нас DCAT» не говорит, примет ли файл data.europa.eu или data.gov.

Профиль	Для кого	Состояние на октябрь 2026 года
DCAT-AP 3.0.1	порталы Европы и data.europa.eu	рекомендация SEMIC от 27 октября 2025 года. Черновик 3.0.2 от 17 июня 2026 года рекомендацией ещё не стал
Национальные профили поверх DCAT-AP	портал одной страны	следуют версии, которую проверяет национальный сборщик. Примеры — DCAT-AP_IT и DCAT-AP-CH
DCAT-US 3.0	федеральные ведомства США, файл `data.json`	опубликована 1 мая 2026 года. Версию 1.1 data.gov продолжает собирать, пока ведомства переходят
DCAT-AP HVD	наборы высокой ценности в ЕС	выпуск 3.0.0, октябрь 2024 года, поверх DCAT-AP
GeoDCAT-AP 3.1.0	перевод карточки ISO 19115 в DCAT	рекомендация SEMIC от 16 февраля 2026 года, приведена к DCAT-AP 3.0.1
StatDCAT-AP	статистические таблицы на портале открытых данных	действующая версия — 1.0.1 (2019). Версия 3.0.0 с 10 июня 2026 года — рабочий черновик
HealthDCAT-AP	каталоги медицинских данных для EHDS	черновик Еврокомиссии; выпуск 7 — май 2026 года. Номер в том же году уже сдвигался, берут последний выпуск
mobilityDCAT-AP 3.0.0	национальные точки доступа к данным о транспорте	рекомендация NAPCORE поверх DCAT-AP 3.0.0, ещё не поверх 3.0.1

Отраслевой профиль не ведут вторым каталогом. Его получают переводом из того описания, которым отрасль уже пользуется. GeoDCAT-AP для того и сделан: геопортал хранит карточку ISO 19115, а портал открытых данных читает её как DCAT. Так же StatDCAT-AP нужен, чтобы таблицы статистической службы появились среди открытых данных, не заменяя SDMX.

В DCAT-AP 3.0.1 у набора обязательны название и описание, а лицензия живёт у способа получения и обязательной не названа. Для человека лицензия, дата обновления и период важнее этого минимума. Полная таблица обязательных и рекомендуемых полей — в главе о полноте метаданных.

Пространства данных публикуют каталог в DCAT и записывают условия использования языком ODRL: кому можно брать данные и для чего. Описание остаётся обычным DCAT, даже когда сам файл закрыт. Как устроен обмен — в главе о других типах каталогов.

Паспорт вместо DCAT

В России описания открытых данных задают не DCAT, а методические рекомендации версии 4.0. Их утвердили протоколом президиума Правительственной комиссии по цифровому развитию от 12 декабря 2024 года № 49пр. Единица описания — паспорт набора: идентификатор, название, описание, владелец, ответственное лицо и его контакты, ссылка на файл, формат, даты публикации, изменения и актуальности, ключевые слова, ссылки на версии данных и структуры.

Машиночитаемый паспорт для разбора чужого сайта — один файл на весь перечень: `meta.csv`, `meta.xml` или `meta.json`. Требования к CSV называют четыре колонки: `property`, `title`, `value` и `format`. Версия рекомендаций записывается полем `standardversion`. Рядом с паспортом лежат файл данных и файл структуры. На сайтах органов власти, сделанных на Gosweb, у каждого набора три CSV: паспорт, структура и данные. Как устроены такие разделы — в главе о порталах открытых данных.

Образец в том же разделе отстал от требований. Эталонный `meta.csv` содержит только колонки `property` и `value` и ссылается на старый адрес стандарта 3.0, `http://opendata.gosmonitor.ru/standard/3.0`, хотя на обложке стоит версия 4.0. Отдельно, в разделе про RDFa, рекомендации советуют размечать HTML словарём schema.org или Dublin Core. Разметка страницы и паспорт в документе не сведены в один профиль.

Практический вывод. Паспорт читает российский портал и его разборщик страниц. data.europa.eu, data.gov и Google Dataset Search его не читают. Если наборы должны находить только внутри этой системы, паспорта достаточно. Если их должен найти кто-то ещё, к паспорту добавляют schema.org на странице и тот профиль DCAT, который назвал конкретный сборщик.

Остальные языки

schema.org/Dataset — разметка в HTML-странице набора, обычно блоком JSON-LD. Её читают поисковики. Google просит название и описание от 50 до 5000 символов; лицензию, автора, период и территорию называет желательными. Разметка описывает один набор за раз и не заменяет файл всего каталога. Страница, которую дорисовывает скрипт уже в браузере, для робота часто пустая. Как это проверить — в главе о находимости.

ISO 19115 описывает пространственные данные: охват, систему координат, качество, ограничения. В файл карточку записывают по ISO 19139 или ISO 19115-3: в XML это видно по префиксам `gmd:` и `mdb:`. INSPIRE говорит, как европейский орган власти обязан отдавать такую карточку. Снимки и другие пространственные файлы описывает STAC: у коллекции тысячи файлов, у каждого свои дата и территория. Подробнее — в главе о геопорталах.

Dublin Core — короткий набор полей: название, автор, дата, тема, права. Его отдают научные репозитории по протоколу OAI-PMH, и его понимают почти все архивы. **DataCite** подробнее и нужен, чтобы зарегистрировать DOI. Действующая схема — 4.7 от 3 марта 2026 года. Обязательны идентификатор, автор, название, издатель, год и тип ресурса. Лицензия в схеме необязательна: её отсутствие не мешает получить DOI и мешает человеку понять, можно ли использовать файлы. Как устроен сам DOI — в главе об идентификаторах.

DDI описывает обследование: вопросник, выборку, переменные, файлы. **SDMX** описывает показатель, его разрезы и ряд. Оба нужны внутри своей отрасли. Наружу, к порталу открытых данных и к поисковику, от них всё равно делают перевод: StatDCAT-AP и schema.org. Разбор — в главах о микроданных и о показателях.

Croissant описывает набор для машинного обучения поверх schema.org: файлы, поля, разбиение на обучающую и тестовую части. Версия 1.1 от 29 января 2026 года добавила происхождение по модели PROV-O, условия использования (DUO и ODRL) и ссылки на внешние словари. В файле версию видно по `dct:conformsTo` со значением `http://mlcommons.org/croissant/1.1`. Площадки в 2026 году ещё отдают 1.0. Карточка для человека — отдельный текст рядом. Подробнее — в главе об ML-каталогах.

SKOS описывает словарь: у понятия есть адрес, названия, код и связи «шире — уже». **ISO/IEC 11179** описывает элемент данных в реестре: определение, допустимые значения, владелец. Это языки каталогов метаданных, а не порталов наборов. Разбор — в главе о каталогах метаданных.

Упаковка и сборка из готовых стандартов

Несколько спецификаций часто попадают в один список со стандартами каталогов, хотя отвечают на другой вопрос: как упаковать файлы или какие готовые стандарты сочетать.

RO-Crate упаковывает файлы исследования и описание к ним. Рядом с данными лежит `ro-crate-metadata.json` — JSON-LD на основе schema.org. Действующая рекомендация — 1.3 от 22 июня 2026 года. Так передают и сдают объект исследования: данные, программы, статью о том, как их получили. Каталог страны этим файлом не собирают. У Dataverse выгрузка RO-Crate есть отдельным экспортёром сообщества, не как формат обмена порталов.

CDIF (Cross-Domain Interoperability Framework) — рекомендации CODATA, какие уже существующие стандарты сочетать, когда данные переходят границу дисциплины. Версия 1.0 вышла в мае 2024 года, версия 1.1 — в июне 2026 года. Для нахождения CDIF советует schema.org и DCAT, а не новый словарь каталога. Свои стандарты дисциплины остаются: DataCite, DDI, ISO 19115.

Data Package и **BagIt** описывают папку с файлами. В первом рядом лежит `datapackage.json` со списком таблиц и словарём колонок — так публикуют набор в Git. Во втором — манифест контрольных сумм: так архив проверяет, что пакет цел. OCFL задаёт, как репозиторий раскладывает такие пакеты у себя на диске. Ни один из трёх не говорит сборщику порталов, как обойти каталог. Про публикацию папкой — в главе о Git-репозиториях, про архивную упаковку — в главе об архивации.

CodeMeta устроен так же, как schema.org для набора, только для программы: авторы, версия, ссылка на репозиторий. Его ставят рядом с данными, если публикуют и код. Каталогом наборов он не становится.

Справочники

Свободный текст в поле «тема» не склеивается между порталами. «Транспорт», «Transport» и «трансп.» для программы — три разные темы. Значение берут из общего списка и записывают его адресом или кодом.

Для открытых данных в Европе это тринадцать тем data theme Бюро публикаций ЕС: сельское хозяйство, экономика и финансы, образование, энергетика, окружающая среда, государство и общественный сектор, здоровье, международные вопросы, правосудие, регионы и города, население и общество, наука и технологии, транспорт. Для геоданных — девятнадцать тематических категорий ISO 19115: биота, границы, рельеф, воды и другие. Свой рубрикатор завести можно. Наружу его сопоставляют с общим списком, как это делают национальные порталы ЕС: внутри своя рубрикация, в DCAT-AP — тема data theme.

Лицензию записывают ссылкой, а не фразой «открытая лицензия». В DCAT-AP ссылка ведёт в словарь лицензий Бюро публикаций ЕС. В репозитории и в файле `CITATION.cff` ту же лицензию узнают по коду SPDX, например `CC-BY-4.0`. Слова «типовые условия использования» российский читатель поймёт, а программа международного каталога — нет: такого кода в её списке нет. Подробнее — в главе о лицензиях.

Язык записывают кодом из словаря языков ЕС или ISO 639, территорию — кодом страны и региона или ссылкой на узнаваемый объект вроде GeoNames, а не словами «наш регион». Код переживает сбор лучше названия: при перекладке из каталога в каталог имя города теряется чаще, чем идентификатор. Издателя в научном описании записывают идентификатором организации из реестра ROR, автора — идентификатором ORCID. Тогда одна организация и один человек не разъезжаются по разным написаниям имени.

Свой справочник оправдан, если общего нет. Тогда его публикуют отдельно, дают каждому значению постоянный адрес и описывают в SKOS. Список в подсказке к полю формы справочником не считается: на него нельзя сослаться из другого каталога.

Как выбрать

Ситуация	Решение
Описания забирает национальный портал или data.europa.eu	профиль и версия, которые проверяет их валидатор. Соседняя версия не подойдёт
Наборы должны находиться поиском	schema.org на странице каждого набора, даже если DCAT уже есть
Нужен DOI	схема DataCite. Запись DCAT DOI не регистрирует
Слой карты должен появиться и на портале открытых данных	внутри ISO 19115, наружу GeoDCAT-AP
Каталог видят только свои сотрудники	внешний стандарт не нужен
Российский орган публикует на федеральный портал и больше никуда	паспорт рекомендаций 4.0
Тот же орган хочет, чтобы наборы нашли снаружи	паспорт оставить и добавить schema.org; DCAT — когда сборщик назовёт профиль
В дисциплине есть своя схема	её хранят внутри, наружу отдают общий профиль

Ситуация	Решение
Хочется завести свои поля	в форме каталога — можно. В файле для чужого сборщика — только известный профиль
Десяток наборов без сборщика	schema.org и один файл со всеми описаниями. Полный DCAT-AP избыточен

Коротко, как выглядит обоснование по шаблону карточки решения. Карточка про то, какое расширение CKAN включить, уже есть в главе об экспорте. Здесь развилка другая: национальный паспорт или «международный стандарт».

- **Контекст.** Министерство публикует около двухсот наборов. Федеральный портал и сайты на Gosweb ждут паспорт по рекомендациям 4.0. Сотрудники просят «сделать как в Европе, в DCAT». Отдельного сборщика, который назвал бы версию DCAT-AP, нет. Наборы хотелось бы видеть в поиске по данным.
- **Варианты.** Заменить паспорт на DCAT-AP. Оставить только паспорт. Вести паспорт для российского портала, разметить страницы schema.org и не писать свой JSON.
- **Критерии.** Кто реально забирает описания, попадает ли набор в поиск, сколько стоит поддерживать два описания одного набора.
- **Решение.** Паспорт оставить. На страницу набора добавить schema.org. DCAT-AP включать тогда, когда конкретный сборщик назовёт версию профиля. Свой формат «почти DCAT» не делать.
- **Почему.** Российский портал читает паспорт. Поисковик читает разметку страницы. DCAT-AP без сборщика никто не проверит, а самодельный JSON придётся объяснять каждому новому адресату отдельно. Два описания берутся из одной карточки: второе не набирают руками.
- **Когда пересмотреть.** Если федеральный портал начнёт собирать DCAT, если набор должен попасть в data.europa.eu или отраслевой агрегатор, либо если DCAT-AP 3.0.2 станет рекомендацией, а черновик перестанут принимать.

Как узнать стандарт по файлу

Команды, которыми проверяют, что выгрузка открывается и проходит валидатор, собраны в главе об экспорте. Здесь — как по содержимому понять, какой язык перед вами. Смотрят начало файла, а не название кнопки «скачать метаданные».

Что видно в файле	Какой это стандарт
Колонки `property` и `value`, поле `standardversion`	паспорт российских методических рекомендаций
`conformsTo` с адресом DCAT-AP, DCAT-US или другой версии профиля	профиль DCAT; номер в адресе важнее слова «DCAT» в инструкции
`@type` со значением Dataset и словарь schema.org в `@context`	разметка schema.org. Если `conformsTo` указывает на `mlcommons.org/croissant`, это Croissant
Файл `ro-crate-metadata.json`	RO-Crate
XML с префиксом `gmd:` или `mdb:`	карточка ISO 19139 или ISO 19115-3
Пространство имён `datacite.org/schema/kernel-4`	схема DataCite

Что видно в файле	Какой это стандарт
`datapackage.json` рядом с таблицами	Data Package, словарь файлов, а не каталог портала

Дальше две проверки. Первая: версия в файле совпадает с той, которую назвал сборщик. DCAT-AP 2 там, где ждут 3.0.1, и DCAT-US 1.1 там, где уже принимают 3.0, — это разные профили. Вторая: значения полей взяты из справочника. Слово «CSV» в поле формата и адрес типа файла из словаря для валидатора не равны, что и показала выгрузка data.gov.gr.

Где обычно ошибаются

Свой JSON называют стандартом. Поля похожи на DCAT, имена свои, справочников нет. Каждый сборщик разбирает такой файл отдельно, и при первой же проверке профиля он отпадает.

Включили выгрузку и считают профиль правильным. У расширения и у облачного сервиса профиль по умолчанию часто старше того, который ждёт страна. Его включают настройкой. На data.gov.gr профиль был включён, а два поля всё равно не прошли валидатор.

Берут черновик как действующую версию. На октябрь 2026 года черновики — DCAT-AP 3.0.2, StatDCAT-AP 3.0.0 и HealthDCAT-AP. Их читают, чтобы понять, куда движется профиль. В обмен отдают ту версию, у которой статус рекомендации и которую принимает валидатор сборщика.

Свободный текст вместо кода. Тема, лицензия, язык и территория, записанные фразой, не склеиваются с чужим каталогом. Обязательное поле, заполненное заглушкой, формально проходит и по смыслу пустое.

Одним стандартом закрывают все задачи. DCAT не регистрирует DOI, DataCite не кормит data.europa.eu, паспорт не попадает в поиск по данным, Croissant не описывает показатель. Каждому адресату — его язык из одной и той же карточки.

Путают описание набора и устройство файла. Выложить закупки «в DCAT» нельзя: DCAT описывает карточку, OCDS — сами контракты. Карточка при этом говорит, что файл соответствует OCDS, через `dct:conformsTo`.

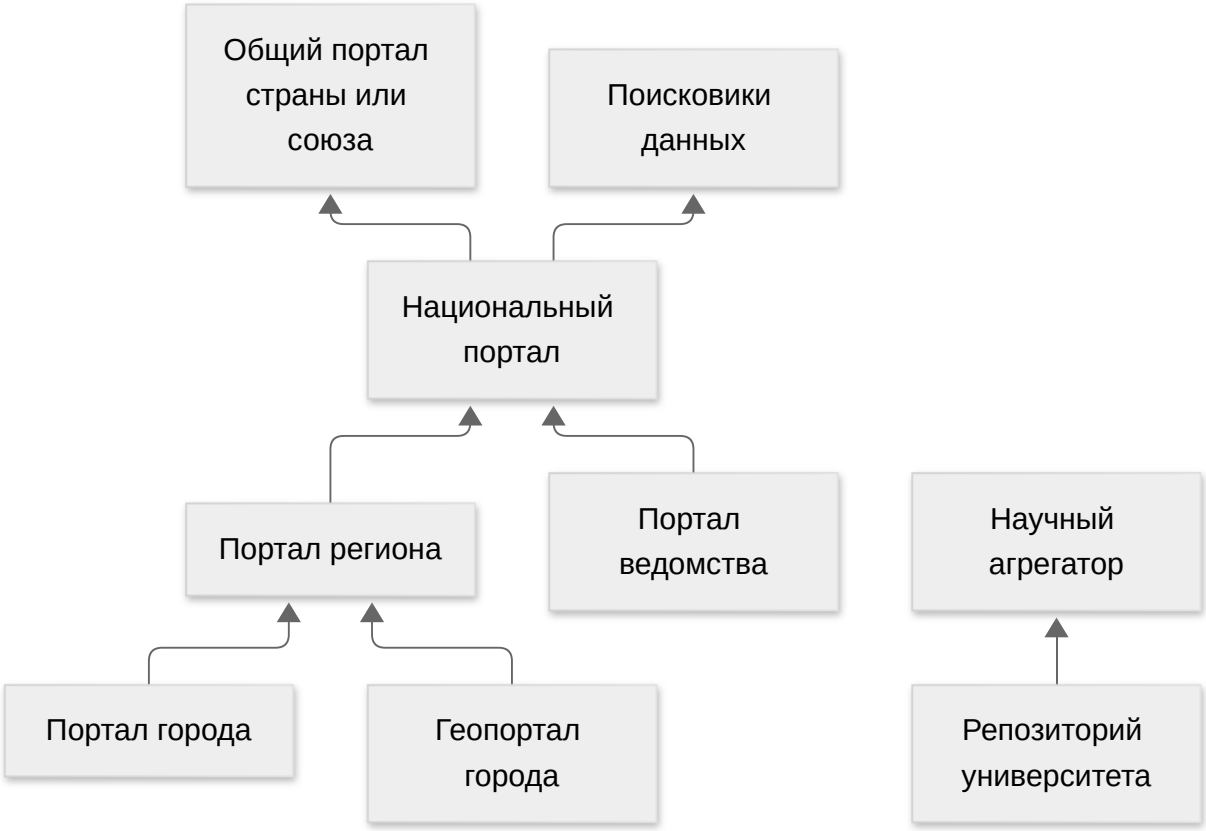
Источники

- DCAT 3, рекомендация W3C от 22 августа 2024 года; предыдущая — DCAT 2 от 4 февраля 2020 года
- DCAT-AP 3.0.1, рекомендация SEMIC от 27 октября 2025 года; черновик 3.0.2 от 17 июня 2026 года
- DCAT-US 3.0, опубликована 1 мая 2026 года; объявление от 26 мая 2026 года. Версию 1.1 data.gov принимает на переходный период
- GeoDCAT-AP 3.1.0, рекомендация от 16 февраля 2026 года; DCAT-AP HVD; StatDCAT-AP 3.0.0, рабочий черновик от 10 июня 2026 года; HealthDCAT-AP; mobilityDCAT-AP 3.0.0
- Схема DataCite 4.7, выпущена 3 марта 2026 года
- Croissant 1.1, опубликована 29 января 2026 года

- RO-Crate 1.3, рекомендация от 22 июня 2026 года
- CDIF, версия 1.1 — июнь 2026 года, текст на cdif.codata.org
- Методические рекомендации по публикации открытых данных, версия 4.0, приложение к протоколу от 12 декабря 2024 года № 49пр; страница документа на сайте Минфина, опубликована 14 января 2026 года. Эталонный `meta.csv` внутри текста ссылается на `http://opendata.gosmonitor.ru/standard/3.0`
- Словарь тем data theme и словарь лицензий Бюро публикаций ЕС
- schema.org/Dataset, разметка Dataset в документации Google
- Каталоги стандартов: FAIRsharing, Data Standards Directory, стандарты на resources.data.gov — проверено 13 октября 2026 года
- И. Бегтин, Стандарты работы с данными, январь 2021 — различие стандартов метаданных и стандартов данных
- Экспорт метаданных, полнота метаданных, порталы открытых данных

Федерации и агрегация

Федерация — цепочка каталогов. Город публикует наборы, регион их забирает, страна забирает у региона, общий портал вроде data.europa.eu забирает у страны. Поисковики данных стоят сбоку и читают тех, кто отдаёт описания машинно.



Забор описаний по расписанию называют сбором, или харвестингом. Берут не файлы данных, а карточки: название, лицензия, адрес, дата. Файл человек или программа скачивает уже у источника.

Что делает источник

- Отдаёт все описания по одному адресу: файл DCAT, API, CSW или OAI-PMH. Можно и постранично, если порядок записей устойчив и до конца списка можно дойти. Поисковая выдача по 10 штук, где порядок меняется от запроса к запросу, для сбора не годится.
- Не закрывает этот адрес капчей и регистрацией.
- Держит адрес набора постоянным. Если набор снят, карточка говорит об этом, а не исчезает без следа.
- Пишет лицензию на наборе. Агрегатор не имеет права её выдумать.

Профиль должен совпадать с тем, который ждёт верхний портал. «Почти DCAT» верхний портал пропускает молча, и город решает, что его данные «не взяли».

Что делает тот, кто собирает

- Ходит по расписанию и не забирает каталог целиком каждые десять минут.
- Хранит адрес источника и дату сбора.
- Не плодит дубли: один набор с городского портала и его копия на региональном — одна запись со ссылкой на первоисточник.
- Убирает или помечает карточки, у которых источник больше не открывается.

У CKAN каркас сбора — расширение ckanext-harvest, сбор RDF — часть ckanext-dcat. У Piveau сбор DCAT — основная работа ядра: платформу Fraunhofer FOKUS построили вокруг сбора и проверки качества описаний, и на ней работает сам data.europa.eu. Подробности расширений — в главе о CKAN.

Как это выглядит в жизни

Европа. Земельный портал Open.NRW собирает описания ведомств и городов Северного Рейна — Вестфалии. Национальный портал Германии GovData забирает Open.NRW вместе с другими землями: в октябре 2026 года у GovData около 10,6 тысячи наборов из Open.NRW. Описания идут в немецком профиле DCAT-AP.de. Общеевропейский data.europa.eu забирает GovData и другие национальные порталы, по умолчанию раз в сутки, и сам переводит названия и описания на все официальные языки ЕС.

США. Каждое федеральное ведомство публикует файл `data.json` по профилю DCAT-US, а сборщик data.gov забирает эти файлы. Состояние каждого источника видно на отдельном сайте harvest.data.gov. После переезда data.gov на собственную разработку в 2026 году сборщик остался отдельной программой, которая пишет в общую базу, а витрина её только читает.

Геоданные ЕС. Страны отдают описания слоёв через каталожные сервисы CSW по правилам INSPIRE. До 1 июля 2026 года их собирал отдельный геопортал INSPIRE, теперь он закрыт, и национальные геокаталоги собирает data.europa.eu, переводя карточки ISO в DCAT по профилю GeoDCAT-AP. Сами слои остаются в национальных сервисах. Подробнее — в главе о геопорталах.

Наука. Репозитории университетов отдают описания по OAI-PMH в формате из руководств OpenAIRE, и OpenAIRE собирает их в общий граф научных результатов.

Таиланд: одна платформа на все уровни. Центральный каталог государственных данных gdcatalog.go.th работает на CKAN, в октябре 2026 года в нём около 25 тысяч наборов. Каталоги провинций — не отдельные сайты на разном ПО, а поддомены центрального на той же платформе, например ranong.gdcatalog.go.th. В центральном каталоге провинции видны как организации, геоданные показывает геопортал geo.gdcatalog.go.th на TerriaJS. Профили не расходятся, потому что источник и сборщик — одна система. Цена — регионы не выбирают себе платформу.

Шотландия: сначала поиск, потом портал. Национального портала у Шотландии долго не было. Вместо него правительство сделало поисковик find.data.gov.scot: он забирает описания с сайтов ведомств, советов и других держателей данных через их API, а где API нет — разбирает страницы. В октябре 2026 года в нём около 26,5 тысячи наборов из 70 источников

от 922 владельцев, у каждого набора своя страница. Только в 2026 году заработал портал data.gov.scot, пока в бета-версии и с 290 наборами: на нём публикуют те, у кого нет своего места, а описания он отдаёт в DCAT и schema.org. Такой порядок полезен, когда данные уже разбросаны по сотне сайтов: поисковик быстро показывает, что есть, а портал нужен для остального.

Протоколы сбора

Протокол	Кто обычно отдаёт	Что получает сборщик
Файл DCAT (RDF, `data.json`)	порталы открытых данных	все описания одним файлом
OAI-PMH	научные репозитории; data.europa.eu предпочитает его и для DCAT-AP	описания постранично, с возможностью забрать только изменённые с прошлого раза; список файлов — только если его содержит сам формат описаний
API CKAN	порталы на CKAN	описания в формате CKAN; data.europa.eu принимает его только от старых систем
CSW	геопорталы	описания слоёв в ISO 19115 / 19139
SPARQL	порталы на графовых базах	ответ на запрос к описаниям
Разметка schema.org и карта сайта	любой сайт	описания со страниц; так читают поисковики данных

Возможность забирать только изменения важна для больших каталогов: сборщик не перекачивает сто тысяч описаний каждую ночь.

Как попасть в сбор

Верхний портал обычно просит заполнить короткую анкету. У data.europa.eu она публичная: адрес точки выгрузки, профиль описаний (DCAT-AP, формат CKAN или INSPIRE), протокол, формат дат по ISO 8601, нужен ли вход. Затем заявку отправляют через форму обратной связи. После подключения портал видит оценку качества своих описаний в MQA. Полезно пройти такую анкету до выбора платформы: она сразу показывает, чего не хватает.

Где цепочка рвётся

Муниципалитет заводит свой портал, который регион не умеет читать. Регион копирует карточки руками. Национальный портал видит регион и не видит города. Поисковик видит три копии одного бюджета.

Чем длиннее цепочка, тем больше теряется по дороге. data.europa.eu — агрегатор агрегаторов: он собирает национальные порталы, которые сами собрали города и ведомства. На нём хорошо видно, где именно рвётся.

Теряются лицензия и источник. В августе 2024 года набор об экологически опасных предприятиях города Мальмё лежал на городском портале под лицензией CC0. Через национальный портал Швеции dataportal.se он доходил до data.europa.eu уже без лицензии и без ссылки на городскую страницу. К октябрю 2026 года город переехал на новый портал, лицензия CC0 теперь доходит, а ссылки на файлы ведут к Мальмё. Но страницы набора на

городском портале в карточке по-прежнему нет: каталогом записан национальный портал Швеции. А старая карточка на data.europa.eu, на которую кто-то мог сослаться, при переезде города исчезла: её адрес отдаёт ошибку 404, потому что агрегатор строит адрес из адреса источника.

Это не единичный случай. По подсчёту И. Бегтина, в 2024 году лицензия была указана примерно у 150–180 тысяч из 1,8 миллиона наборов data.europa.eu, то есть не больше чем у 10%. В октябре 2026 года лицензия в виде ссылки на известный текст есть примерно у 775 тысяч наборов — меньше половины из 1,73 миллиона, которые в том же месяце показывал портал. Стало лучше, но больше половины наборов всё ещё доходят без машиночитаемой лицензии, а заметная часть «лицензий» — значения вроде «условия неизвестны».

Вывод для каждого звена: передавать лицензию и ссылку на страницу набора у источника (в DCAT это свойство `dcat:landingPage`) и не менять идентификатор набора при переезде.

Один источник перевешивает страну. В октябре 2026 года на data.europa.eu около 1,73 миллиона наборов, из них 850 тысяч — немецкие. 681 тысяча приходит из одного каталога — геокаталога GDI-DE, и почти 450 тысяч из них — планы застройки отдельных участков (Bebauungsplan), каждый как отдельный набор, обычно в виде сервисов WMS и WFS. Национальный портал GovData даёт ещё 169 тысяч. В 2024 году картина была та же: 609 из 726 тысяч немецких наборов шли из GDI-DE. Поэтому «в Германии половина данных Европы» говорит не об открытости, а о том, как геокаталог режет данные. Сравнивая страны в агрегаторе, смотрите на число источников и типы данных, а не на общий счёт.

Видна только часть страны. В июне 2024 года на data.europa.eu было 324 набора из Эстонии: портал собирал только национальный геопортал. К октябрю 2026 года подключили и национальный портал открытых данных, и наборов стало 8,7 тысячи, 8,3 тысячи из них — с этого портала. Сколько данных страны дошло наверх, зависит от того, какие каталоги она зарегистрировала. Проверить это можно фильтром «каталог» в поиске верхнего портала: если своего каталога там нет, его наборов нет тоже.

Протокол не передаёт файлы. OAI-PMH передаёт только описания: что представляет собой сам ресурс, спецификация прямо выводит за свои рамки. Обязательный формат описаний — простой Dublin Core, где нет поля для списка файлов. Поэтому научные агрегаторы, которые собирают репозитории по OAI-PMH, обычно хранят ссылку на страницу набора, а не на файлы. Репозиторию, который хочет, чтобы его данные можно было скачать прямо из поиска, стоит рядом с OAI-PMH отдавать DCAT или API со списком файлов.

Свой формат сбора. Российский data.gov.ru собирал перечни наборов, которые ведомства вели по собственной схеме из методических рекомендаций. Ни один зарубежный агрегатор этот формат не читает, поэтому цепочка обрывалась на первом же шаге за пределами страны. Подробнее — в антипаттернах.

Честнее не заводить нижний портал, если верхний уже принимает публикации организаций. Это обычный вывод для небольшого города, см. пример муниципалитета.

Источники

- Экспорт метаданных

- API и массовая выгрузка
- ckanext-harvest, ckanext-dcat
- Piveau и проект data.europa.eu у Fraunhofer FOKUS
- Как попасть в сбор data.europa.eu — анкета, протоколы, ежедневный сбор, машинный перевод
- GovData и профиль DCAT-AP.de; число наборов Open.NRW — запрос к API GovData, 11 октября 2026 года
- DCAT-US, harvest.data.gov, приложение каталога data.gov
- data.europa.eu, INSPIRE datasets move to the European Data Portal, 22.06.2026 — закрытие геопортала INSPIRE 1 июля 2026 года; руководства OpenAIRE
- data.europa.eu: число наборов по странам, каталогам, ключевым словам и лицензиям — API поиска; наборы с лицензией-ссылкой — запрос к точке SPARQL; карточка набора Мальмё и старый адрес карточки (404) — API поиска; всё 12 октября 2026 года
- Каталог государственных данных Таиланда (CKAN 2.9.2, API `package_search`), каталог провинции Ранонг, геопортал, 12 октября 2026 года
- find.data.gov.scot — счётчики на главной; data.gov.scot, бета-версия, 12 октября 2026 года
- Спецификация OAI-PMH 2.0 — природа ресурса вне рамок протокола, обязательный формат Dublin Core
- И. Бегтин, Telegram: data.europa.eu как агрегатор агрегаторов: Мальмё и лицензии, 04.08.2024; GovData и геокаталог GDI-DE, 20.08.2024; Эстония в европейском портале, 12.06.2024; Таиланд, 15.08.2024; find.data.gov.scot, 03.11.2024; почему не собирать OAI-PMH, 02.04.2024

Лицензии и условия доступа

Лицензия отвечает на вопрос, что можно делать с данными. Условия доступа отвечают на другой: кто может забрать файл. Это два поля. Фраза в подвале сайта не заполняет ни одно из них.

Коротко:

- На каждый открытый набор ставят одну известную лицензию: ссылкой на текст и тем кодом, который ждёт сборщик. На карточке, у файла и в выгрузке стоит одно и то же.
- Для обычных открытых данных это CC BY 4.0 или CC0. Свой текст на две страницы и значение «не указана» сборщик не узнает.
- Если закон страны требует свой документ, его оставляют для человека, а в поле для программы пишут известную лицензию, которая этому документу не противоречит.
- Закрытый файл получает условия доступа и срок ответа. Открытую лицензию на него не ставят: она обещает то, чего файл не отдаёт.

Три условия, при которых данные называют открытыми, — в главе о порталах открытых данных. Кто видит карточку и кто скачивает файл — в главе о ролях и доступе.

Четыре пометки, которые путают

Пометка	На какой вопрос отвечает	Куда её пишут
Лицензия	что можно делать с данными: копировать, менять, использовать в работе, в том числе коммерческой	поле набора и каждого файла, та же ссылка в выгрузке
Условия доступа	кто получает файл и что для этого нужно сделать	поле доступа на карточке: открыто, после регистрации, по заявке, только в защищённой среде
Правила сайта	как вести себя на портале: не перегружать сервер, не выдавать себя за владельца	страница «Условия использования», не поле набора
Указание источника и ссылка в статье	как назвать автора	лицензия может требовать указать источник; научная ссылка устроена отдельно, см. главу о цитировании

Противоречие, которое встречается часто: на наборе стоит CC BY, а в правилах сайта написано «только некоммерческое использование». Тогда непонятно, какой текст побеждает. Оставляют один.

Что считать открытой лицензией

Открытое определение (версия 2.1) называет открытым то, что любой может свободно получить, использовать, изменять и передавать дальше. Указать источник при этом можно требовать. Потребовать, чтобы производная база осталась на тех же условиях, тоже можно. Запрет коммерческого использования, запрет изменений и формула «спросите разрешения» открытыми данными это определение не считает.

Отсюда практический список.

Лицензия	Что разрешает	Когда её берут
CC0 1.0	отказ от прав, насколько это позволяет закон; указывать автора не требуют	справочники и факты, которые должны без трения попадать в другие базы
CC BY 4.0	копировать, менять, в том числе в коммерческих целях, указав автора	обычный выбор для открытых данных
CC BY-SA 4.0	то же, но производное произведение остаётся на тех же условиях	когда важно, чтобы переработки оставались открытыми, и набор не собираются смешивать с чужими базами
ODbL 1.0	лицензия для базы данных; производную базу отдают на тех же условиях	базы, которые ведут как базу. Так опубликована карта OpenStreetMap
Национальная открытая лицензия	по смыслу близко к указанию автора	когда так уже принято в стране и текст есть в списке сборщика

Версию пишут явно: «4.0», а не «CC BY». В тексте CC BY 4.0 отдельно разобрано особое право на базу данных, которое признают в Европейском союзе. Регламент о наборах высокой ценности тоже называет эту версию. Статья 4(3) регламента (ЕС) 2023/138 требует отдавать такие наборы на условиях CC0, либо CC BY 4.0, либо равноценной или более свободной открытой лицензии. CC BY-SA и ODbL строже: у производной базы должен остаться тот же текст. Для наборов высокой ценности они не подходят. Две лицензии с сохранением условий ещё и плохо смешиваются между собой: у одной базы не может быть сразу двух таких текстов. Если цель — чтобы набор брали в чужие базы, берут CC BY 4.0 или CC0.

CC BY-NC и CC BY-ND в этот список не входят. Первая запрещает коммерческое использование, вторая запрещает изменения.

Какую ставить

Ситуация	Лицензия
Справочник, который должны копировать в другие базы без лишних условий	CC0 1.0
Обычный открытый набор органа власти, науки или компании	CC BY 4.0
Набор высокой ценности по регламенту ЕС 2023/138	CC0 или CC BY 4.0, либо национальная лицензия не строже них
Нужно, чтобы переработки оставались открытыми, и смешивать базу ни с кем не планируют	CC BY-SA 4.0 или, для базы, ODbL
В стране уже есть своя открытая лицензия, и сборщик её узнаёт	эта лицензия
Закон требует национальный текст, у которого нет кода в списке сборщика	текст для человека и рядом известная лицензия, которая ему не противоречит
Файл закрыт	условия доступа, не открытая лицензия
Рядом с данными лежит программа	у программы своя лицензия, у таблицы своя

Национальные тексты, которые сборщики уже знают: Open Government Licence v3.0 в Великобритании, Licence Ouverte 2.0 во Франции, Data licence Germany — attribution — Version 2.0 в Германии. У каждой есть код SPDX и код в словаре лицензий Бюро публикаций ЕС.

Таблица кодов — ниже.

Типовые условия в России

Российские порталы часто ссылаются на «Типовые условия использования общедоступной информации, размещаемой в сети Интернет в форме открытых данных». Их утвердили протоколом заочного голосования Правительственной комиссии по координации деятельности открытого правительства от 19 сентября 2016 года № 6.

В тексте, который открылся 14 октября 2026 года, три пункта тянут в разные стороны. Пункт 10 просит использовать данные в законных целях, давать ссылку на источник и не искажать сведения. Пункт 13 прямо разрешает и коммерческое, и некоммерческое использование. Пункт 12 говорит, что договор не нужен и специализированные соглашения и лицензии не требуются. Международного кода у документа нет. В паспорте методических рекомендаций 4.0 поля лицензии тоже нет: паспорт читает российский разборщик, а сборщик за пределами страны это поле в паспорте не найдёт, даже если его туда добавить своими силами.

Практический вывод. Страницу с типовыми условиями оставляют для человека. В поле, которое уходит в выгрузку, юрист выбирает известную лицензию, которая этим пунктам не противоречит. Ближе всего к «ссылка на источник и коммерческое использование разрешено» оказывается CC BY 4.0. Если требовать ссылку на источник не хотят, ближе CC0. Сам документ этого соответствия не объявляет: его должен подтвердить юрист органа, а не карточка каталога.

Где она стоит

Лицензия стоит на каждом наборе. Общая лицензия на весь каталог допустима только когда она и правда одна и скопирована в выгрузку каждого набора. Подвал сайта сборщик не читает.

В DCAT-AP 3.0.1 лицензия — рекомендуемое свойство способа получения (файла или сервиса), не набора. Обязательной профиль её не называет. Профиль сознательно не вешает лицензию на набор: если у набора один текст, а у файла другой, непонятно, какой побеждает. Для человека и для сборщика поле всё равно нужно. Где оно стоит среди других свойств — в главе о полноте метаданных.

У CKAN лицензия хранится на наборе. В выгрузку DCAT её переносит расширение ckanext-dcat. По документации выпуска 2.4.4 файл наследует лицензию набора только если включена настройка `ckanext.dcat.resource.inherit.license`. По умолчанию она выключена. Заполненное поле набора при этом может не попасть к файлу в RDF.

14 октября 2026 года на data.gov.gr у набора `emy-station-116` записи одной лицензии разошлись. В API набора код `cc-by`, а ссылка ведёт на страницу Open Definition без номера версии. У первого файла ссылка уже на понятие `CC_BY_4_0` словаря EC. В RDF, который открылся по адресу `dataset_series/emy-station-116.rdf`, у самого набора свойства лицензии нет: `CC_BY_4_0` стоит у способа получения и у сервиса данных. Кто проверит только `license_url` в API, версии 4.0 не увидит. Кто проверит только поле набора в выгрузке DCAT,

решит, что лицензии нет, хотя у файла она есть. На data.gov.gr в выборке из главы о полноте метаданных лицензия была у 32% наборов и у 99% файлов. Смотреть нужно оба места и выгрузку.

Сборщик лицензию ещё и теряет по дороге. На data.europa.eu в октябре 2026 года ссылка на известный текст была примерно у 40% записей о наборах. Как это происходит — в главе о федерациях.

Отдельно стоит пустышка. Значение «лицензия не указана» — тоже значение. На data.gouv.fr встроенная оценка засчитывает его как заполненную лицензию. Код CKAN `other-open` («какая-то другая открытая лицензия») говорит, что лицензия вроде бы открытая, и не говорит, какой у неё текст. Строгая проверка ждёт ссылку на известный текст. Как считают доли — в главах о полноте и о метриках.

Три записи одной лицензии

Одну и ту же лицензию программы ищут по-разному. Файл `CITATION.cff` и список SPDX ждут код вроде `CC-BY-4.0`. CKAN по умолчанию ждёт свой короткий код без версии. DCAT-AP рекомендует ссылку на понятие словаря ЕС. Zenodo берёт список SPDX, но в API пишет код строчными буквами: у записи 10.5281/zenodo.4898406 14 октября 2026 года в поле стояло `cc-by-nc-4.0`.

Лицензия	SPDX	Код CKAN по умолчанию	Код словаря ЕС
CC0 1.0	`CC0-1.0`	`cc-zero`	`CC0`
CC BY 4.0	`CC-BY-4.0`	`cc-by`	`CC_BY_4_0`
CC BY-SA 4.0	`CC-BY-SA-4.0`	`cc-by-sa`	`CC_BYSA_4_0`
ODbL 1.0	`ODbL-1.0`	`odc-odbl`	`ODC_BL`
Open Government Licence v3.0	`OGL-UK-3.0`	`uk-ogl`	`OGL_3_0`
Licence Ouverte 2.0	`etalab-2.0`	в списке по умолчанию нет	`ETALAB_2_0`
Data licence Germany, attribution, 2.0	`DL-DE-BY-2.0`	в списке по умолчанию нет	`DLDE_BY_2_0`

Ссылка словаря ЕС собирается так: `http://publications.europa.eu/resource/authority/licence/` и код из таблицы. Код ODbL в этом словаре — `ODC_BL`. По буквам «ODbL» его не найти.

В ядре CKAN 2.12 пятнадцать лицензий. У `cc-by` и `cc-zero` ссылка ведёт на страницу Open Definition, а не на текст Creative Commons, и версии 4.0 в коде нет. У `uk-ogl` ссылка как раз на текст версии 3.0. У `other-open` ссылки нет. Свой список подключают настройкой `licenses_group_url`: это файл JSON, а не правка ядра. 14 октября 2026 года data.gov.gr отдавал те самые пятнадцать значений. open.canada.ca список заменил: у национальной лицензии код `ca-ogl-lgo` и ссылка на текст Open Government Licence — Canada. В списке SPDX она записана как `OGL-Canada-2.0`.

У data.gouv.fr коды тоже свои. 14 октября 2026 года набор обязательных данных о закупках региона Овернь — Рона — Альпы в API имел код `fr-lo`. В RDF у способа получения стояла ссылка на PDF 2014 года, `etalab.gouv.fr/wp-content/uploads/2014/05/Licence_Ouverte.pdf`.

В тот же день адрес открылся и перенаправил на PDF с именем licence-ouverte-2-0. Кода `ETALAB_2_0` в ссылке нет. Сборщик, который узнаёт только словарь ЕС, эту лицензию не сопоставит, хотя по ссылке лежит нужный текст.

Несколько прав в одной посылке

У таблицы, у программы рядом с ней и у фотографий в ней могут быть разные права. В карточке пишут лицензию данных. Лицензию программы кладут в файл `LICENSE` кода. Если внутри набора есть чужие строки, это сказано в описании: чужая лицензия не становится лицензией всего набора от того, что файл лежит в одной папке.

Четыре места должны называть один текст: страница, файл `CITATION.cff`, `datapackage.json` и запись на общем хабе. Расхождение уже разобрано на примере Global Jukebox: в метке репозитория файл `LICENSE.txt` — CC BY 4.0, а запись Zenodo по DOI — CC BY-NC 4.0. Подробности — в главе «Когда портал не нужен».

Платформы добавляют ещё одну путаницу: согласие публикатора с архивом и лицензия для того, кто скачает данные.

В форме сдачи DSpace 10 обязательный шаг «License» — согласие с архивом. Текст лежит в `default.license` и может быть своим у коллекции. Шаг «CC License» — отдельный и необязательный: им задают лицензию самой записи. Галочка «я согласен сдать работу в архив» лицензией набора не является.

В Dataverse на запись ставят лицензию из списка установки или, если установка это разрешает, свои условия. Оба сразу документация не предлагает. Часто, если человек ничего не выбрал, подставляется CC0: это настройка установки. Нормы сообщества просят всё равно сослаться на набор. Документация прямо говорит, что эти нормы не заменяют лицензию и не создают юридической обязанности.

Когда файл закрыт

Ограниченный доступ нужен для персональных, чувствительных и лицензируемых данных. Описание при этом публикуют открыто: набор можно найти и запросить. Закрывать карточку вместе с файлом — прятать само существование данных. Уровни, платформы, которые умеют оставить описание открытым, и свойство `dct:accessRights` разобраны в главе о ролях и доступе.

Здесь важное дополнение. Открытая лицензия на закрытом файле врёт. CC BY обещает, что файл можно копировать. Если файл выдают по заявке, на карточке пишут условия доступа, порядок запроса и срок ответа. «По запросу» без срока — не уровень доступа.

Регистрация тоже барьер, даже если в поле стоит CC BY. Портал, который отдаёт данные только после проверки личности, по трём условиям открытости остаётся общедоступным. Так устроены портал провинции Гуандун и, до отмены регистрации, национальный портал Казахстана: примеры — в главах о порталах открытых данных и об антипаттернах.

Персональные данные перед публикацией обезличивают. Оценка, можно ли по оставшимся полям узнать человека, — отдельная работа. Если риск остаётся, файл не кладут в открытый каталог.

По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) ограниченный доступ отмечен примерно у 1,1 тысячи каталогов. Это пометка каталога целиком, не доля закрытых наборов внутри него.

Насколько лицензию видно снаружи

Там же, в выгрузке реестра, текст или ссылку на лицензию на уровне всего каталога удалось заполнить примерно у 1% записей (около 420 из 43 тысяч). Это значит, что лицензию трудно найти даже тому, кто описывает каталоги. Модель «лицензия на весь каталог» или «лицензия у каждого набора» определена чаще, примерно у 17 тысяч каталогов, поровну на оба варианта. Ещё у 5,6 тысячи модель явно неизвестна. Остальные записи сюда не относятся: модель у них не названа. Эти три числа не складываются в 43 тысячи. Текст лицензии и модель лицензии — разные поля. Запросы — в приложении.

Долю лицензий внутри портала по этой пометке не считают. Её считают по выгрузке наборов, как в главах о полноте и о метриках.

Пример: свой текст или известная лицензия

Контекст. Региональный орган публикует около двухсот открытых наборов и несколько обследований с персональными данными. Юрист принёс двухстраничные «условия использования» и просит поставить их на всё. Национальный портал забирает описания в DCAT-AP и узнаёт лицензию по словарю EC. Часть наборов входит в перечень, для которого ждут лицензию не строже CC BY 4.0.

Варианты. Оставить текст юриста в подвале. Вписать этот текст в поле лицензии каждого набора. На открытые файлы поставить CC BY 4.0 ссылкой словаря EC, страницу юриста оставить для человека, если она не противоречит CC BY, а закрытые обследования описать условиями доступа.

Критерии. Сборщик узнаёт лицензию без разбора абзаца. Текст на карточке и текст в выгрузке не спорят. Закрытый файл не обещает свободное копирование.

Решение. Открытые наборы — CC BY 4.0, в выгрузке ссылка `.../licence/CC_BY_4_0`, в `CITATION.cff` код `CC-BY-4.0`. Страница с национальным текстом остаётся, юрист письменно подтверждает, что она не ужесточает CC BY. Обследования — открытое описание и условия доступа со сроком ответа, без CC BY.

Почему. Подвал в выгрузку не попадает. Свой абзац не совпадает ни с одним кодом словаря, и национальный портал покажет набор без лицензии. Одна пометка на открытые и закрытые файлы врёт про одну из двух групп. CC BY-SA отпадает: она строже, чем допускает регламент для наборов высокой ценности, и мешает смешивать базы.

Последствия. На CKAN включают наследование лицензии файлом или заполняют поле файла отдельно и проверяют RDF, а не только `license_url`. Список лицензий ядра, где у CC BY нет версии, заменяют своим файлом. У обследований появляется срок ответа, которого в старых «условиях» не было.

Когда пересмотреть. Страна опубликует национальную лицензию с кодом SPDX, который словарь EC уже содержит. Регламент или национальный портал назовёт другой допустимый текст. Обследование обезличат до файла, который можно отдавать открыто.

Как записать свою карточку — в главе «Как обосновать решение».

Как сделано в платформах

Платформа	Где живёт лицензия	Что важно
CKAN 2.12, ядро	поле `license_id` набора, список из пятнадцати значений	у CC BY нет версии, ссылка ведёт на Open Definition; свой список — файл по настройке `licenses_group_url`
ckanext-dcat 2.4.4	переносит лицензию в DCAT	файл получает лицензию набора только при `ckanext.dcat.resource.inherit.license`; по умолчанию выключено
udata, ядро	обязательное поле из своего списка	значение «не указана» существует, и балл карточки data.gouv.fr его засчитывает
Dataverse, ядро	лицензия из списка установки или свои условия	часто по умолчанию CC0; нормы цитирования лицензией не являются
Zenodo на InvenioRDM	обязательное поле, список на основе SPDX, можно несколько и свой текст	в API код строчными буквами; подсказка по умолчанию — CC BY 4.0
DSpace 10, ядро	согласие с архивом обязательно; лицензия записи — отдельный шаг Creative Commons	согласие сдать работу в архив в `dc.rights` само не попадает
GeoNetwork	правовые ограничения карточки ISO 19115	короткого кода вроде `cc-by` в ядре нет; в DCAT ссылка попадёт, если профиль выгрузки переложил это поле

Уровни доступа тех же платформ — в главе о ролях и доступе. Там же сказано, кто умеет оставить описание открытым при закрытом файле.

Публикация без портала

Рядом с файлами кладут одну лицензию в четыре места, и текст везде один:

- `CITATION.cff` — код SPDX, в шаблоне книги это `CC-BY-4.0`;
- `datapackage.json` — тот же код в `licenses[].name` и ссылка на текст в `licenses[].path`;
- разметка schema.org — ссылка на текст в свойстве `license`;
- запись Zenodo или другого хаба — лицензия, которую хаб принимает своим кодом.

Шаблоны лежат в каталоге заготовок. GitHub узнаёт лицензию по файлу `LICENSE` в корне, если текст совпадает с известным. Файл в корне не заменяет поле записи на хабе: в примере Global Jukebox они разошлись.

Как проверить

Откройте три случайные карточки, не самые верхние в списке.

- Лицензия есть на карточке.
- Это ссылка на известный текст или код из таблицы выше. Абзац «данные можно использовать» и значение «не указана» не считаются.
- В выгрузке у способа получения та же ссылка. Если в DCAT поля нет, сборщик лицензию потеряет, даже если на сайте она написана.
- Если у набора и у файла стоят разные ссылки, это записано специально, а не потому, что список платформы не знает версию.
- Если файл закрыт, карточка открывается без входа, открытой лицензии на ней нет, и написано, как запросить файл и в какой срок ответят.

Для портала на CKAN команда показывает код, ссылку набора и лицензию первого файла. Подставьте свой идентификатор.

```
curl -s "https://data.gov.gr/api/3/action/package_show?id=emy-station-116" \
  | jq '{id:.result.license_id, url:.result.license_url,
         file:(.result.resources[0].license // "пусто")}'
```

14 октября 2026 года ответ был таким: код `cc-by`, ссылка `http://www.opendefinition.org/licenses/cc-by`, у файла — понятие `CC_BY_4_0`. Список, из которого взялся код:

```
curl -s "https://data.gov.gr/api/3/action/license_list" \
  | jq '[.result[] | {id, url}]'
```

В RDF ту же ссылку ищут у способа получения. Адрес `dataset/emy-station-116.rdf` в тот день открыл `dataset_series/emy-station-116.rdf`, и там стояло `licence/CC_BY_4_0`.

Источники

- Open Definition 2.1
- Creative Commons, CC0 1.0, CC BY 4.0 и текст лицензии, CC BY-SA 4.0
- Open Data Commons Open Database License
- Open Government Licence v3.0, Licence Ouverte 2.0, Data licence Germany — attribution — Version 2.0
- Словарь лицензий ЕС, запрос к точке SPARQL Бюро публикаций 14 октября 2026 года; коды SPDX
- DCAT-AP 3.0.1, раздел о правовой информации: лицензия у способа получения, рекомендация словаря ЕС и CC BY 4.0 для данных FAIR
- Регламент (ЕС) 2023/138, статья 4(3)
- Типовые условия использования общедоступной информации в форме открытых данных, протокол от 19 сентября 2016 года № 6. Текст пунктов 10, 12 и 13 прочитан 14 октября 2026 года на странице территориального управления Россельхознадзора, которая

воспроизводит документ. Файл PDF на data.nalog.ru в тот же день открылся; в свойствах файла стоит 22 сентября 2016 года. Страница документа на сайте Минфина из этой сети не открылась

- Список лицензий CKAN 2.12: ckan/model/license.py. Наследование в выгрузке: ckanext-dcat 2.4.4, ключ `ckanext.dcat.resource.inherit.license`, значение по умолчанию `false`
- Выбор лицензии в Dataverse 6.12, лицензии Zenodo, форма сдачи DSpace 10
- Проверка 14 октября 2026 года: API и RDF data.gov.gr (набор `emy-station-116`, `license_list` из 15 значений), API data.gouv.fr (набор закупок региона Овернь — Рона — Альпы, код `fr-lo`, в RDF ссылка на PDF 2014 года), `license_list` open.canada.ca (код `ca-ogl-lgo`), API Zenodo записи `4898406`
- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года
- Полнота метаданных, роли и доступ

Находимость

Набор опубликован, когда его можно найти, не зная адреса портала наизусть. Иначе каталог — внутренний диск с красивой обложкой.

Эта глава — про каталог целиком: что платформа должна делать сама для сотен и тысяч страниц, где регистрировать каталог и как проверить, что поисковики его видят. Как сделать находимой одну страницу или репозиторий без портала, написано в главе «Как сделать публикацию находимой».

Поисковики

Google Dataset Search и похожие системы читают разметку schema.org/Dataset на странице набора. Рядом нужна карта сайта sitemap, по которой поисковик обходит все страницы наборов: по выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) это самый частый машинный адрес каталогов, около 7 тысяч записей. Название и описание пишут человеческим языком. Набор «Выгрузка 12» не ищется.

На портале разметку и карту сайта не пишут руками: их должна выдавать платформа для каждой страницы.

Платформа	Разметка schema.org на странице набора	Карта сайта
CKAN	расширение ckanext-dcat, модуль structured data	не в ядре; сторонние расширения ckanext-sitemap, сообщество CKAN их не ведёт, самое свежее обновлялось в 2024 году
Dataverse	в ядре	в ядре: создаётся командой администратора, большой каталог делится на несколько файлов
InvenioRDM	в ядре	в ядре начиная с v13: создаётся и обновляется сама, указана в `robots.txt`
GeoNetwork	в ядре: JSON-LD schema.org на странице записи; поля Croissant для схемы ISO 19115-3 добавлены в ветку будущего выпуска 4.4.13, в 4.4.12 их нет	в ядре, но `robots.txt` со ссылкой на неё лежит внутри адреса GeoNetwork, а не в корне сайта; карту отправляют в поисковики вручную
Облачные сервисы	проверьте на странице набора: разметка бывает урезанной	проверьте адрес `/sitemap.xml`

Google советует ставить разметку на одну главную страницу набора, а на копиях — например, в результатах поиска или на портале, который собрал набор у источника, — ссылаться на неё свойством `sameAs`. Для федерации это важно: если региональный портал повторяет карточку города с полной разметкой и без ссылки на оригинал, поисковик видит два разных набора.

Что поисковик не видит

Поисковый робот читает HTML-страницу и идёт по ссылкам. Всё, до чего нельзя дойти так, для него не существует, даже если человек в браузере видит это без труда.

Интерфейс, который рисует скрипт. Страница приходит пустой оболочкой, а список наборов и описания дозагружает скрипт из API. Робот, который не выполняет скрипты, видит пустую страницу; веб-архив — тоже. Так устроен раздел «Сборники» на сайте Росстата: в HTML страницы нет ни одной ссылки на документ, список подгружается отдельным запросом (так он выглядел в веб-архиве в феврале 2026 года). Документы Росстата всё равно находятся в поиске, но в основном потому, что на них ссылаются другие сайты. Так же сделана и страница набора на data.europa.eu: в исходном HTML нет ни описания, ни разметки, всё дорисовывает скрипт (октябрь 2026 года). Архив бюллетеней Счётной палаты отдаёт документы через поисковый API, и до самих PDF роботы не доходят — этот случай разобран в главе об архивации.

Лечится это не отказом от API, а тем, что у каждого набора есть обычная страница: сервер сразу отдаёт HTML с описанием, ссылками на файлы и разметкой, а списки наборов листаются обычными ссылками. API — в дополнение к страницам, а не вместо них.

Сайт закрыт для зарубежных адресов. Поисковики, веб-архивы, агрегаторы и ИИ-агенты ходят из облачных дата-центров в разных странах. Если сайт не отвечает на запросы из-за рубежа, он выпадает отовсюду сразу. 12 октября 2026 года с зарубежного сервера не открылись ни rosstat.gov.ru, ни ach.gov.ru. Что это значит для каталога и как проверить, разобрано в антипаттерне «Дверь закрыта для программ» и в главе об архивации.

Карта сайта, о которой поисковик не знает. Поисковик ищет `robots.txt` только в корне сайта. GeoNetwork кладёт свой `robots.txt` со ссылкой на карту сайта внутрь своего адреса, например `/geonetwork/srv/api/robots.txt`, и документация прямо советует регистрировать карту в поисковиках вручную. В октябре 2026 года у geodata.nz, geocat.ch и эстонского каталога геоданных файла `robots.txt` в корне не было. Разметку schema.org на страницах записей GeoNetwork 4.4 отдаёт, но без карты сайта робот до этих страниц может и не добраться. Поэтому каталоги на GeoNetwork находят в основном через агрегаторы — data.europa.eu, который с 1 июля 2026 года заменил геопортал INSPIRE, и национальные геокаталоги, — а не через обычный поиск.

Разметка есть и у спама

Разметку schema.org/Dataset ставит сам владелец страницы, и никто её не проверяет. Проект Web Data Commons извлекает такую разметку из открытого веб-индекса Common Crawl. В выпуске 2024 года тип Dataset нашёлся на 694 тысячах страниц примерно 2 тысяч сайтов. По ручной оценке И. Бегтина (июнь 2024 года), из 1,4 миллиона размеченных так объектов примерно треть — мусор (например, обычные статьи справочных сайтов, размеченные как наборы данных), треть — страницы продавцов данных и только треть — данные, которые можно скачать.

Тип DataCatalog, которым размечают каталог целиком, встречается так редко, что Web Data Commons не выделяет его в отдельную выборку ни в 2023, ни в 2024 году. По проверке реестра каталогов в 2024 году, такая разметка была примерно у одного каталога из тысячи.

Крупные порталы тоже размечают по-разному. В 2024 году data.gov отдавал урезанную разметку — название и ключевые слова. В октябре 2026 года на новом каталоге catalog.data.gov в разметке набора есть и лицензия, и список файлов. У data.europa.eu

разметки в HTML нет вовсе.

Отсюда практический вывод: разметка сама по себе не отличает ваш набор от спама. Заполните в ней то, чего у спама обычно нет: лицензию, издателя, список файлов с прямыми ссылками (`distribution` с `contentUrl`) и ссылку на каталог (`includedInDataCatalog`), как в примере в документации Google.

Агрегаторы и реестры

Верхний портал найдёт вас только по выгрузке, о которой вы договорились. Как это устроено и как попасть в сбор, разобрано в главе «Федерации и агрегация».

Независимые агрегаторы и поисковики данных — Dateno, find.data.gov.scot, научные агрегаторы — тоже не обходят весь интернет. Они ведут список источников и для каждого типа платформы держат свой сборщик. Отсюда два надёжных пути попасть к ним:

- **Публиковать там, где уже собирают.** Национальный или ведомственный портал, Zenodo, отраслевой репозиторий: их сборщики уже настроены, и новый набор попадёт в агрегаторы без ваших усилий.
- **Ставить типовое ПО с известным интерфейсом.** CKAN отдаёт свой API и DCAT, DKAN — файл `data.json`, JKAN — общий `data.json`, InvenioRDM — REST API и OAI-PMH. Сборщики для них уже написаны, достаточно сообщить адрес каталога. Самописному порталу со своим интерфейсом нужен отдельный сборщик, и до него агрегатор доходит последним или не доходит вовсе.

Насколько это важно, видно на российском примере. В 2024 году из 154 российских каталогов в реестре Dateno проиндексировали около 15. Причины назвали такие: государственные сайты закрыты для зарубежных адресов, перечни по методическим рекомендациям неудобно собирать, а для 40–50 самописных порталов нужен свой сборщик для каждого. Финляндию, где каталоги работают на типовом ПО, собрать было гораздо проще.

Дополнительно каталог регистрируют там, где его аудитория ищет сам перечень порталов:

- национальный реестр или верхний портал страны;
- re3data для научного репозитория — заявку подаёт сам репозиторий через форму на сайте, после проверки появляется карточка с тематикой, условиями и сертификатами;
- FAIRsharing для базы данных или стандарта;
- реестр dataportals-registry — он описывает каталоги для тех, кто ищет портал, а не набор.

Регистрация не заменяет выгрузку. Карточка в реестре со сломанным адресом хуже, чем отсутствие карточки: после переезда портала обновите и её.

Программы

Программе нужны стабильный API или файл всех описаний и документация без капчи. Файл `llms.txt` — короткая записка для языковой модели о том, что есть на сайте; это предложение формата, а не принятый стандарт. MCP — способ отдать программе список запросов каталога. И то и другое имеет смысл, когда каталогом уже пользуются такие программы. Они не чинят отсутствие DCAT. Подробнее в главе о трендах.

Как проверить, что каталог находят

Подставьте адрес портала и адрес страницы одного набора:

```
PORTAL=https://portal.example
PAGE=https://portal.example/dataset/budget-2026

# карта сайта отвечает и указана в robots.txt
curl -s "$PORTAL/robots.txt" | grep -i sitemap
curl -s -o /dev/null -w "%{http_code}\n" "$PORTAL/sitemap.xml"

# на странице набора есть разметка Dataset
curl -s "$PAGE" | grep -o '"@type": *"Dataset"' | head -1
```

Дальше три шага, которые командой не проверить:

1. Прогоните страницу набора через проверку расширенных результатов Google — она покажет ошибки в разметке Dataset.
2. Добавьте сайт в Google Search Console и отправьте карту сайта. Там видно, сколько страниц поисковик проиндексировал и почему пропустил остальные.
3. Через несколько недель найдите пять-десять наборов по названию в Google Dataset Search. Если находятся копии с агрегаторов, а не ваши страницы, проверьте `sameAs` и то, что страница набора открывается без входа.

Короткий порядок

Адрес, schema.org, карта сайта и выгрузка в нужном профиле уже стоят в чек-листе требований. Здесь к ним добавляются регистрация каталога в профильном реестре, отправка карты сайта в Search Console и проверка через месяц: набор находится поиском, выгрузка открывается.

Источники

- Разметка Dataset в документации Google — главная страница набора, `sameAs`, карта сайта
- ckanext-dcat: Google Dataset Search; сторонние ckanext-sitemap — поиск по GitHub, 11 октября 2026 года
- Карта сайта в Dataverse, экспорт schema.org в Dataverse
- Карты сайта в InvenioRDM, выпуск v13
- GeoNetwork 4.4 и поисковики — `robots.txt` внутри адреса каталога, совет регистрировать карту сайта вручную; RFC 9309 — `robots.txt` читается только из корня сайта; отсутствие `robots.txt` в корне у geodata.nz, geocat.ch и metadata.geoportaal.ee — проверка 12 октября 2026 года
- Раздел «Сборники» Росстата — снимок в веб-архиве от 13 февраля 2026 года (список загружается запросом к `/compendium/getPage`); недоступность rosstat.gov.ru и ach.gov.ru с зарубежного сервера и HTML страницы набора на data.europa.eu — проверка 12 октября 2026 года
- Web Data Commons: выборки schema.org, выпуск 2024 года и выпуск 2023 года
- Разметка набора на catalog.data.gov — проверка страницы набора 12 октября 2026 года

- JKAN, DKAN, InvenioRDM
- И. Бегтин, Telegram: Росстат и Ajax, 16.11.2022; архив Счётной палаты, 30.11.2022; GeoNetwork и поисковики, 01.04.2023; качество разметки Dataset в Common Crawl, 20.06.2024; Dataset и DataCatalog в schema.org, 15.09.2024; как попасть в индекс Dateno, 22.10.2024; российские каталоги в Dateno, 15.08.2024; геоблокировки и ИИ-агенты, 14.07.2025
- Регистрация репозитория в re3data
- llms.txt, Model Context Protocol
- Экспорт метаданных
- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. Карта сайта отмечена примерно у 7 тысяч каталогов.

ЧАСТЬ VIII.

Публикация без портала

Статические сайты, Git-репозитории, облачные хранилища и общие хабы: когда их достаточно и как сделать такую публикацию находимой.

1. Когда портал не нужен
2. Статические сайты
3. Git-репозитории
4. Облачные хранилища и общие хабы
5. Как сделать публикацию находимой

Когда портал не нужен

Свой портал нужен, когда без него нельзя выполнить обязанность или когда наборов и публикаторов уже столько, что страница и файл описаний перестают справляться. До этого данные публикуют в чужом каталоге, на своём сайте, в Git-репозитории или на общем хабе. Так делают и небольшие группы, и федеральные ведомства: национальный каталог США забирает с сайта ведомства файл описаний, а файлы данных оставляет на месте.

Пустая установка обещает каталог, сопровождать который некому. Описания устаревают. Если регистрация открыта, она собирает спам — так бывает у порталов на CKAN, подробнее в главе о доступе и ролях. После окончания договора на хостинг пропадают адреса. Страница с десятком наборов такого обещания не даёт.

Что в этой главе называется порталом

Каталог — любое место, где набор можно найти, понять и получить. Им бывает и статический сайт. Портал в этой главе — своё приложение: база описаний, поиск, учётные записи, роли и сервер, который кто-то обновляет годами. Список наборов — страница или файл со ссылками, без карточек и поиска. Для десятка наборов список нормален. Тупиком он становится, когда у строки нет лицензии, даты и постоянного адреса.

Как каталог собран из частей — в главе об устройстве каталога. Как довести список до находимой публикации — в главе о списках наборов.

Пять вопросов до установки

Ответы записывают до выбора платформы. Платформа, которая закрывает функции, бесполезна, если её некому сопровождать. Сравнение способов создания — в главе «Свои силы, свободное ПО или облачный сервис».

1. **Кто должен увидеть данные.** Вышестоящий каталог, поисковик, читатель статьи, коллеги внутри организации. От адресата зависит файл описаний: DCAT-AP, DCAT-US, schema.org или DataCite.
2. **Сколько наборов и как часто они меняются.** Десятки наборов и правки несколько раз в год закрывает страница или репозиторий. Сотни наборов, которые каждый день приходят из учётной системы, требуют загрузки в портал или скрипта, который сам обновляет файлы. Скрипт — по-прежнему публикация без портала.
3. **Сколько организаций публикует и как они правят.** Одна группа обходится правами на репозиторий и проверкой изменений перед слиянием. Портал нужен, когда публикаторов много и они не работают с Git, либо когда поток заявок на закрытый файл при открытом описании уже не разобрать почтой и ручным одобрением на общем хабе.
4. **Кто обновляет сервер через три года.** Если ответ «автор страницы» или «никто», портал ставить рано. Ставка или договор на сопровождение делают портал возможным, но первые три вопроса не отменяют.

5. **Что именно обязан забрать верхний каталог.** Обычно это один файл описаний в нужном профиле. API поиска нужен людям и программам, когда наборов уже сотни. Сборщику хватает файла.

Ответ	Куда публиковать
Одна группа, десятки наборов, верхний каталог принимает чужие организации	в этот каталог, своей витрины нет
Свой сайт уже есть, верхний каталог забирает файл	файл описаний на этом сайте
Нужны история правок и совместная работа	Git-репозиторий и страница из тех же файлов
На данные будут ссылаться в статье, своего репозитория нет	общий репозиторий с DOI, например Zenodo
Файл большой	облачное хранилище или хаб, описание — на своей странице
Много публикаторов вне Git, поток заявок на закрытые файлы, сбор чужих каталогов	свой портал или типовой портал, который уже раздают органам власти

Последняя строка не значит «ставить CKAN завтра». Типовой портал, который регион или поставщик уже ведёт для муниципалитетов, закрывает роли без вашей установки. Разбор для города — в примере муниципалитета.

Куда публиковать

Пять способов. Их сочетают: репозиторий хранит таблицы, сайт показывает их людям, Zenodo выдаёт DOI на выпуск.

Способ	Когда его хватает	Чего не будет	Подробнее
Чужой каталог	наборов мало, каталог уже принимает ваши наборы	своего домена и своих полей	способ создания
Раздел своего сайта	сайт есть, наборов немного	формы «добавить набор» и ролей для внешних публикаторов	ниже, на примере файла `data.json`
Статический сайт	десятки наборов, есть человек, который правит файлы	поиска по API и сбора чужих каталогов по расписанию, пока его не напишете отдельно	статические сайты
Git-репозиторий	важны история и совместная правка	больших файлов в самом Git и находимости без страницы	Git-репозитории
Облако и общий хаб	файл большой, нужен DOI или аудитория уже на площадке	гарантии, что правила площадки сохранятся	хранилища и хабы

Минимум, без которого ни один из способов не работает: постоянный адрес, лицензия, дата и контакт, разметка schema.org. Файл со всеми описаниями добавляют, когда наборов несколько или их забирает сборщик: с этого момента публикация становится каталогом, и для неё действует общий минимум. Как его собрать по шагам — в главе «Как сделать публикацию находимой». Какие возможности карта считает обязательными и как закрыть их без портала — в карте возможностей.

Файл на своём сайте вместо портала

data.gov не хранит файлы ведомств. Ведомство публикует каталог метаданных — обычно `data.json` по профилю DCAT-US — по постоянному адресу на своём сайте. Сборщик забирает этот файл по расписанию, и записи появляются в общем каталоге. В инструкции GSA источник — публичный файл, тип источника «документ», частота от ежедневной до ежемесячной. По правилам сборщика источник забирают не реже раза в месяц.

13 октября 2026 года список источников сбора показывал 793 записи. 135 из них имеют тип «документ»: сборщик забирает один файл. У 117 адрес ведёт на `data.json`. Ещё 577 принадлежат Бюро переписи: это папки географических описаний, нарезанные по типам границ, а не отдельные ведомства. Для небольшого ведомства вывод прямой: ему достаточно одного файла.

Живой пример — Избирательная комиссия США (Election Assistance Commission). Её сайт собран на Drupal 10. Каталог данных комиссии — файл eac.gov/data.json. 13 октября 2026 года в нём было 11 записей по схеме DCAT-US 1.1, и сборщик data.gov забирает его раз в неделю как источник `eac-data`. Этого хватает, чтобы описания попали в национальный каталог.

Сбор забирает и слабое описание. Поля `license` в записях нет. В поле `rights` схема DCAT-US 1.1 ждёт короткий текст об условиях доступа, а у части записей стоит логическое значение `true`, у части поля нет. Часть адресов скачивания ведёт на раздел сайта, а не на файл. Часть записей — отчёты в PDF. Файл вместо портала закрывает задачу сбора. Лицензию, прямую ссылку и отделение набора от отчёта автор проставляет сам.

В Европе тот же приём: data.europa.eu забирает каталог по DCAT-AP, если издатель подал его на сбор. Свой CKAN для этого не требуется. Как устроена цепочка — в главе о федерациях.

Три места сразу

Исследовательский проект Global Jukebox публикует данные о песенных традициях без своего портала. Сайт theglobaljukebox.org показывает записи людям. Таблицы и код лежат в репозиториях организации theglobaljukebox на GitHub. Выпуск набора Cantometrics, на который ссылаются, заархивирован в Zenodo: DOI 10.5281/zenodo.4898406, версия v0.1-alpha от 3 июня 2021 года. 13 октября 2026 года сайт открывался, запись Zenodo тоже.

Сочетание работает, пока у трёх мест одно и то же описание. Здесь оно разошлось уже в том выпуске, который получил DOI. Файл `LICENSE.txt` в метке v0.1-alpha — лицензия CC BY 4.0. Файл `.zenodo.json` той же метки указывает CC BY-NC 4.0, и запись Zenodo показывает именно её. Человек, который цитирует DOI, и человек, который скачал репозиторий, получают разные условия. Лицензию выбирают один раз, пишут в том месте, которое считают главным, и в остальные копируют ту же.

Пример: ведомство с двадцатью наборами

Контекст. Небольшое ведомство. Около двадцати наборов: выгрузки раз в квартал и годовые отчёты. Сайт на обычной CMS уже есть, его ведёт один веб-мастер. Национальный каталог забирает файл описаний в известном профиле — DCAT-US или DCAT-AP — и организации внутри себя тоже принимает. Своей команды порталов нет.

Варианты. Свой CKAN. Облачный портал. Организация только на национальном каталоге, файлы остаются на сайте ведомства. Файл `data.json` на существующем сайте, зарегистрированный для сбора.

Критерии. Попасть в национальный каталог. У каждого набора — лицензия и прямая ссылка на файл. Адрес файла описаний не меняется при правке текста на сайте. Сопровождение укладывается в работу одного веб-мастера. Своя витрина на уже существующем сайте важнее отдельного портала: читатели ищут данные там, где ищут само ведомство.

Решение. Файл описаний на существующем сайте и заявка на сбор. Если национальный каталог внешние файлы не забирает и принимает только организации внутри себя — публиковать туда и портал не ставить.

Почему. Свой CKAN при одном веб-мастере останется без обновлений ядра: критерий сопровождения он проигрывает. Облачный портал закроет поиск и карточки, но плата и перенос адресов не окупаются на двадцати наборах. Организация только на национальном каталоге тоже попадает в сбор и даже проще в сопровождении, но свою витрину ведомство при этом теряет. Файл на своём сайте оставляет витрину там, где она уже есть, и отдаёт сборщику тот профиль, который он ждёт.

Последствия. Ведомство зависит от профиля и расписания сбора. Ошибка в файле видна после прогона сборщика, поэтому файл проверяют валидатором до заявки. Своего поиска по наборам нет: при двадцати записях хватает страницы со ссылками.

Когда пересмотреть. Наборов стало больше двухсот, появилось несколько подразделений с разными правами, национальный каталог перестал принимать файл, или часть файлов нужно закрыть, оставив описание открытым, и заявок стало больше, чем разбирает один человек.

Как записать такую карточку у себя — в главе «Как обосновать решение».

Когда свой портал всё-таки нужен

- Наборы публикуют много организаций, эти люди не работают с Git, и каждой нужны свои права и проверка перед публикацией.
- Описание открыто, а файл выдают по заявке, и заявок столько, что общий хаб и почта не справляются. Так устроены каталоги микроданных.
- Нужно регулярно забирать чужие каталоги и модерировать копии. Скрипт в репозитории, как у opendata.scot, тянет тысячи описаний, но роли он не даёт.
- Наборы каждый день приходят из учётной системы, и скрипт, который обновляет файлы, поддерживать труднее, чем загрузку в портал.
- Профиль страны требует возможности, которой нет ни у файла, ни у чужого каталога, уже принимающего ваши наборы. Чаще требуют именно файл в нужном профиле.

Для отдельных типов каталога порог другой. Научной группе с одним набором к статье свой репозиторий не нужен. Городу с тридцатью наборами — тоже.

Задача	Где разобрано, когда свой каталог не нужен
Открытые данные органа власти	порталы открытых данных

Задача	Где разобрано, когда свой каталог не нужен
Карты и слои	геопорталы
Данные к статье	научные репозитории
Показатели	каталоги показателей
Микроданные	каталоги микроданных
Наборы для моделей	ML-каталоги
Справочники	каталоги метаданных

Чем платите

Способ	Что получаете	Чем платите
Чужой каталог	карточки, сбор и поиск уже работают	чужие поля, чужое расписание, нет своего домена
Раздел сайта	дёшево, адрес в вашем домене	правки вручную; поиск по API не появляется
Статический сайт	почти нечему ломаться, сайт можно забрать файлами	правки через файлы; форму «добавить набор» люди без Git не осилят
Git-репозиторий	история и совместная работа	предел размера: GitHub не принимает файлы больше 100 МиБ и советует держать репозиторий меньше 1 ГБ; без страницы с разметкой набор плохо ищется
Облако и хаб	большие объёмы, DOI, чужая аудитория	правила и лимиты площадки меняются, площадка может закрыться

Цены чужих площадок сравнивают в главе о способе создания. История закрытых хабов — в главе об облачных хранилищах.

Что видно по живым спискам

В реестре dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) 107 каталогов записаны как списки наборов: страница, репозиторий или листинг хранилища, без карточек. Программный интерфейс есть у 15. Карта сайта — у 12. 95 записаны своей разработкой. Остальные двенадцать — страницы на обычных движках сайтов: Drupal, WordPress, TYPO3, Jekyll, SharePoint и Liferay. Недействующих и снятых — 7.

Малая доля интерфейсов для списка ожидаема. Списку из двадцати ссылок API поиска не нужен. Ломается другое: у строки нет лицензии, адрес страницы меняется при переделке сайта, разметки schema.org нет, и ссылки умирают молча. Как это измерили на тематическом списке и как починить — в главе о списках наборов. Реестр лицензии и разметку schema.org почти не записывает, поэтому эти дыры в его цифрах не видны.

Как проверить своё решение

- [ ] Назван адресат: кто забирает описания и в каком профиле.
- [ ] Посчитаны наборы, организации-публикаторы и частота обновления.

- [ ] Назван человек, который будет обновлять публикацию через три года. Если это сервер портала — назван и тот, кто обновляет сервер.
- [ ] У каждого набора есть лицензия, дата и прямая ссылка на файл, а не только страница раздела.
- [ ] Файл со всеми описаниями открывается без входа и проходит проверку того профиля, который ждёт сборщик.
- [ ] Записано событие, при котором решение открывают заново.

Команды проверки файла и разметки — в главах об экспорте метаданных и о находимости.

Типичные ошибки

Портал «как у всех». Город или ведомство ставит CKAN, потому что так сделал сосед. Через два года ядро не обновлено, наборов столько же, сколько влезло бы на одну страницу. Сначала отвечают на пять вопросов выше.

Ссылка на папку. Общий диск или бак без страницы описания — не публикация. Адрес временный, лицензии нет, через год папку не находят.

Свой формат файла описаний. Чтобы собрать сайты, придумывают таблицу CSV со своими колонками. Сборщик национального каталога её не читает. Файл делают в том профиле, который уже принимает верхний каталог. Как это вышло с методическими рекомендациями в России — в главе об антипаттернах.

Ждать, пока кто-то построит портал. Пока национальный или региональный каталог не принимает наборы, страница и репозиторий уже публикуют данные. Сообщество в Шотландии так собрало opendata.scot, не дожидаясь государственного портала. Подробности — в главе о статических сайтах.

Три копии — три лицензии. Сайт, репозиторий и хаб описывают один набор по-разному. Читатель не знает, какая копия главная. Главной назначают одну и сверяют с ней остальные.

Источники

- GSA, How to get your Open Data on Data.gov — data.gov не хранит файлы, ведомство публикует `data.json`; Harvester API — источник забирают не реже раза в месяц; список источников сбора — 793 источника, из них 135 с типом «документ», 117 адресов `data.json` и 577 источников Бюро переписи, проверено 13 октября 2026 года
- eac.gov/data.json — 11 записей, схема DCAT-US 1.1, поле `rights` и адреса скачивания; сайт комиссии отвечает заголовком Drupal 10. Проверено 13 октября 2026 года. Источник `eac-data` в списке сбора — еженедельный документ. Поле `rights` в схеме набора DCAT-US 1.1 — текст до 255 знаков или пустое значение; `license` — необязательная ссылка
- Global Jukebox: сайт, организация на GitHub, DOI 10.5281/zenodo.4898406 (запись Zenodo, лицензия CC BY-NC 4.0, версия v0.1-alpha от 3 июня 2021 года). В той же метке репозитория LICENSE.txt — CC BY 4.0, а .zenodo.json указывает CC BY-NC 4.0. Проверено 13 октября 2026 года

- Большие файлы на GitHub — предел 100 МиБ на файл и рекомендация держать репозиторий меньше 1 ГБ
- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года: 107 списков наборов, интерфейс у 15, карта сайта у 12, своя разработка у 95, недействующие и снятые — 7. Запросы 15 октября 2026 года — в приложении
- Как попасть в сбор data.europa.eu

Статические сайты

Статический сайт — это готовые страницы. Генератор собирает их из файлов, чаще всего из Git-репозитория, а хостинг только отдаёт результат: GitHub Pages, GitLab Pages и другие. Базы и круглосуточного приложения нет. Уже собранный сайт переживает паузу в разработке генератора, а забрать каталог к себе — значит скопировать файлы.

Этого хватает, когда наборы публикует небольшая группа, а читателю и сборщику нужны страница набора и один файл описаний. Так устроен opendata.scot: 14 октября 2026 года в его файле описаний было 2596 наборов. Роли, очередь заявок на закрытый файл и API поиска — повод сменить способ публикации. Сервер, поставленный рядом с генератором, этих задач не решает: получится самодельный портал, который некому сопровождать.

Порядок «репозиторий, разметка, DOI, запись в чужом каталоге» — в главе «Как сделать публикацию находимой». Здесь — из чего собирают сам сайт и где у готовых инструментов кончается поставка.

Когда его хватает

Ситуация	Куда
Десятки или несколько тысяч открытых наборов, правят файлами один-двое	статический каталог: JKAN или свой генератор
Наборов мало, сайт ведомства уже есть, сборщику нужен один файл	файл описаний на этом сайте. См. «Когда портал не нужен»
Показатели целей устойчивого развития, людям нужна витрина	OpenSDG
Снимки и другие пространственные файлы, их читает программа	статический каталог STAC в хранилище. См. геопорталы и облако
Одна большая таблица, по ней нужны страница и запрос	Datasette. Это уже маленький сервер, не папка файлов
Много публикаторов без Git, закрытые файлы, API поиска	портал или чужой каталог

Как он собирается

Исходники лежат в Git: по файлу на набор, шаблон страницы, лицензии. Генератор на каждом изменении собирает HTML, общий файл описаний и `sitemap.xml`. Хостинг кладёт эту папку на адрес сайта. Читатель получает обычные страницы. Сборщик забирает один файл, а не ходит по кнопкам.

Сбор чужих каталогов в такой схеме возможен, но это скрипт рядом с сайтом, а не функция генератора. Скрипт по расписанию скачивает описания, кладёт их файлами в репозиторий, генератор собирает страницы заново. Так сделан opendata.scot. Готовый сайт по-прежнему ничего не обходит и никого не модерирует: в нём лежат те файлы, которые скрипт успел положить.

На GitHub Pages у этой схемы есть потолок. Документация на 14 октября 2026 года говорит: опубликованный сайт не больше 1 ГБ, сборка обрывается после 10 минут, у репозитория-источника рекомендуемый предел тоже 1 ГБ, мягкий предел трафика — 100 ГБ в месяц. На opendata.scot файлы данных лежат у публикаторов, а в репозитории — описания, поэтому сайт в этот предел помещается. Каталог, в который положили сами таблицы, упрётся в гигабайт раньше, чем в число наборов. У другого хостинга пределы свои; перед публикацией их читают у этого хостинга.

Страница должна содержать текст описания в HTML. Сайт, который рисует карточку только скриптом в браузере, для сборщика и поисковика пуст. Как это проверяют — в главе о находимости портала.

Чем собирать

Инструмент	Когда его берут	Чего в поставке нет
JKAN	открытые наборы, от десятков до нескольких тысяч	API поиска, ролей, schema.org; файл описаний короче схемы, которую он называет
Обычный генератор (Hugo, Jekyll, Astro)	наборов мало, шаблон сайта уже есть	готового каталога: страницу, разметку и файл описаний пишут сами
PortalJS	свой интерфейс поверх файлов или уже существующего каталога	ролей и API поиска в свободной сборке из файлов
OpenSDG	витрина показателей целей устойчивого развития	живого сервиса SDMX и общего каталога произвольных файлов
Статический каталог STAC	пространственные файлы, которые читает программа	ответа на запрос «все файлы этой территории за эти даты»
Datasette	одна база SQLite	статики: страница жива, пока запущен процесс

В реестре dataportals-registry выпуска 1.27.0 от 15 октября 2026 года — 48 порталов открытых данных на JKAN, 72 каталога на OpenSDG и 44 портала открытых данных на Datasette. Это те же числа, что в главах о порталах открытых данных, показателях и в карточке JKAN.

Обычному генератору хватает шаблона страницы, блока schema.org из готового файла и файла `data.json` или `datapackage.json` рядом. Один набор в Git подробнее разобран в главе о репозиториях. Генератор берут, когда наборов мало и JKAN со своей схемой только мешает. Когда наборов десятки, готовый каталог экономит шаблон.

JKAN

JKAN — каталог на Jekyll, лицензия MIT. Набор — файл Markdown в каталоге `_datasets`. Адрес страницы берётся из пути этого файла: правка заголовка адрес не меняет, переименование файла меняет. Рядом собираются страница организации и общий `data.json`.

Файл объявляет схему DCAT-US 1.1 строкой `conformsTo`. Шаблон заполняет фиксированный список полей: название, описание, категорию, контакт, лицензию и ссылки на файлы. Поле, добавленное в схему сайта, видно на странице. В `data.json` оно попадает, только если его отдельно вписали в шаблон этого файла. Схема DCAT-US 1.1 требует у набора ещё ключевые слова, дату изменения, издателя, идентификатор, уровень доступа и, для федеральных органов США, коды бюро и программы. Шаблон их не выводит. Издателя он пытается взять и

теряет: в шаблоне на ветке `main` переменная, куда издателя записали, и переменная, откуда его читают, названы по-разному. Строка `conformsTo` при этом ставится всегда. Сборщику, которому нужен DCAT-AP, этот файл не отдают, пока шаблон не переделают. Так же сказано в сценарии муниципалитета.

Разметки schema.org в шаблоне страницы нет. Есть RDFa: у блока страницы стоит `dcat:Dataset`. Документация Google говорит, что читает и такой DCAT на странице, но проверка расширенных результатов пример с DCAT не принимает. В шаблон добавляют schema.org из файла книги: эту разметку тест проверяет до индексации. Как устроены поля — в главе о находимости без портала.

Поиск на сайте — фильтр в браузере по файлу `datasets.json`. Отдельного API у этого поиска нет. Общий файл для сборщика — `data.json`.

Форма редактора есть: это Decap CMS, страница `/editor`. Человек входит учётной записью GitHub, правка становится коммитом. Ролей и очереди заявок форма не даёт. На Netlify вход настраивается у хостинга. На GitHub Pages инструкция JKAN просит отдельный маленький сервер входа. Это уже процесс, который кто-то должен держать: ради формы на бесплатных страницах GitHub статика перестаёт быть только файлами. Кто готов править Markdown, этот сервер не включает.

Выпуск с номером — 2.0.0 от 30 марта 2023 года. Следующего номера к 14 октября 2026 года нет. Ветка `main` жила и после выпуска. Карта сайта, плагин `jekyll-sitemap`, добавлена 29 июля 2026 года и в тег 2.0.0 не входит. Последний push ветки — 7 сентября 2026 года. В 2018–2020 годах коммиты были, в основном по оформлению. В 2021 году их не было. В конце 2022 года встроенный редактор заменили на Decap. Уже собранный сайт от такой паузы не падает. Устаревает шаблон: копия с тега 2.0.0 карты сайта не получит, пока её не обновят.

PortalJS

PortalJS — средства на Next.js, лицензия MIT. Свободная сборка по документации на 14 октября 2026 года по умолчанию читает CSV и JSON из файлов проекта и умеет выгрузить статический сайт в папку. Тот же интерфейс на этапе сборки может заранее забрать страницы из уже работающего CKAN. Пока сборка статическая, хостинг снова отдаёт файлы, а новые наборы появляются после следующей сборки. Режим, в котором страница каждый раз спрашивает сервер, требует хост с Node, а не папку файлов.

Платный хостинг — отдельный продукт. Тарифы на тот же день: 1188, 3588 и 14 990 долларов в год. Роли есть уже на первом платном тарифе. Поиск по API и управляемый CKAN — со второго, за 3588 долларов в год. Это портал с подпиской, а не способ обойтись файлами. Сравнение цен — в главе о способе создания. Схема «готовый каталог и свой сайт поверх него» — там же, в разделе о готовом ядре.

OpenSDG

OpenSDG — статический сайт показателей целей устойчивого развития, лицензия MIT. Выпуск 2.4.0 вышел 19 августа 2025 года. Данные и тексты показателей лежат в отдельном репозитории, программа сборки готовит из них файлы, Jekyll собирает страницы. Так

устроены показатели ЦУР Великобритании: 14 октября 2026 года в коде страницы был плагин `jekyll-open-sdg-plugins`.

SDMX здесь — источник, который сборка может прочитать, и файлы SDMX-ML, которые она может положить на сайт, если это включили. Живого сервиса SDMX, к которому обращается Евростат или МВФ, сайт не поднимает. Национальной службе для обмена нужен PxWeb или .Stat, а OpenSDG — для публичной витрины. Выбор расписан в главе о каталогах показателей.

Статический каталог STAC

Статический каталог STAC — дерево JSON-файлов. В корне лежит `catalog.json`, дальше — коллекции и объекты со ссылками на снимки, растры и другие пространственные файлы. Рекомендация спецификации прямая: это файлы на веб-сервере, часто в объектном хранилище; каталог можно обойти по ссылкам, на запрос он не отвечает. Имена файлов — `catalog.json` и `collection.json`. Ссылки внутри дерева делают относительными, если каталог должны суметь перенести.

Пока каталог читает программа, отдельный сервер каталога не нужен. STAC Browser открывает и такое дерево, и STAC API. Запрос «все объекты в этом прямоугольнике за эти даты» дерево само не выполняет: для него файлы загружают в API. Живой пример дерева — публичный каталог Google Earth Engine, файл `catalog.json`. 14 октября 2026 года это был документ типа `Catalog` с 133 дочерними ссылками, 25 841 байт. Рядом есть HTML-страница тех же наборов; JSON нужен программе.

Datasette

Datasette публикует базу SQLite: веб-страница и JSON API появляются одной командой. Лицензия Apache-2.0. Это сосед статики, а не она сама: процесс должен работать, иначе страницы нет. Сопровождение проще портала и всё же это сопровождение процесса.

Рядом есть Datasette Lite: та же программа в браузере, через WebAssembly. Хостинг отдаёт страницу и файл базы, читатель загружает базу к себе. Так публикуют одну таблицу, которую реально открыть целиком. Каталог из тысяч файлов в браузер не кладут.

Два живых каталога

Два работающих примера показывают, что статический каталог годится не только для небольшой коллекции: один из них — национальный портал, другой собран сообществом без государственного бюджета.

Национальный портал Чехии

У data.gov.cz на одном адресе два разных изделия. Информационная часть — статьи, инструкции, стандарты — по репозиторию сайта собрана Jekyll и опубликована через GitHub Pages. README на 14 октября 2026 года говорит то же. Правку текста предлагают запросом на изменение, как в любом открытом проекте.

Каталог наборов — отдельное приложение. README отсылает к нему явно: интерфейс национального каталога живёт в другом репозитории, LinkedPipes DCAT-AP Viewer. Описания лежат в хранилище связанных данных, к хранилищу открыта точка запросов SPARQL.

Связанные данные — способ публиковать описания так, чтобы у записи и у значения был свой адрес и на них можно было сослаться из другой системы.

Файлы наборов портал не хранит. 14 октября 2026 года SPARQL насчитал 31 272 набора из 133 каталогов и 39 624 разных адреса скачивания. На сам data.gov.cz из этих адресов вели 16. Это собственные материалы каталога: списки школ, библиотек и спортплощадок, выгрузка каталога, отчёты о качестве ссылок. Остальные адреса — сайты публикаторов. Плата за такую лёгкость в том, что портал не сохранит копию, если ведомство удалит файл. Почему это важно — в главе о хранилище.

Число наборов меняется сильнее, чем устройство. 24 февраля 2023 года их было около 142 тысяч, из них около 137 тысяч давала одна кадастровая служба. К октябрю 2026 года в точке SPARQL осталось 31 272. Каталог по-прежнему указатель, а не архив.

Каталог, который собрало сообщество

В Шотландии долго не было общего государственного портала открытых данных, хотя свои порталы были у городов, ведомств и университетов. Волонтёры сделали opendata.scot: сайт на JKAN, описания и сборщики лежат на GitHub. Скрипты забирают описания с порталов на CKAN, ArcGIS и других платформах и складывают их файлами в репозиторий. С июля по 9 октября 2026 года синхронизация шла примерно раз в неделю. Последний такой коммит — 9 октября, в ту же минуту обновлён файл на сайте.

В `data.json` в тот день, 14 октября, было 2596 наборов. Файл весит 5 878 536 байт и объявляет схему DCAT-US 1.1. У каждой записи шесть полей: название, описание, категория, контакт, лицензия, ссылки. Издателя нет ни у одной. Ключевых слов, даты изменения, идентификатора и уровня доступа тоже нет. На странице набора организация при этом есть: у блока стоит издатель в RDFa. Сборщик, который читает только `data.json`, её не увидит.

Лицензия в файле заполнена и не всегда пригодна. У 1407 из 2596 это ссылка на Open Government Licence 3.0. У 462 в поле записано `No licence`: ячейка не пустая, пользоваться данными по ней нельзя. У 232 наборов нет ни одной ссылки на файл. Сами файлы лежат на сайтах городов и ведомств. Среди самых частых адресов — порталы на ArcGIS Hub.

Страница набора, открытая в тот же день, не содержит schema.org. В ней есть `dcat:Dataset`, как в шаблоне JKAN. Карта сайта открывается. Поиск на сайте читает `datasets.json` размером 5 775 288 байт. Отдельного API поиска нет.

На сайте есть раздел аналитики: какие организации публикуют данные и на каких платформах работают их порталы. Хостинг страниц — GitHub Pages. Сервер каталога не нужен, пока устраивают файл описаний и страницы. Позже правительство Шотландии сделало свой поисковик по данным. Он разобран в главе о поисковиках и агрегаторах.

Что сделать на сайте

- Адрес страницы набора не меняется при правке опечатки. В JKAN для этого не переименовывают файл.
- В HTML страницы есть schema.org/Dataset. RDFa со словом `dcat:Dataset` это не заменяет.

- Один файл со всеми описаниями открывается без входа. В записи есть издатель, лицензия, дата изменения и прямая ссылка на файл. Строку `conformsTo` сверяют с полями.
- В поле лицензии — ссылка на известную лицензию или явный текст условий. Значение `No licence` лицензией не считается.
- Есть `sitemap.xml`.
- Текст описания лежит в HTML, а не появляется только после выполнения скрипта.

Без этих пунктов сайт остаётся списком ссылок. С ними он закрывает минимум из чек-листа. Как довести одну публикацию до поисковика и чужого каталога — в главе о находимости.

Когда статического сайта уже мало

Пора менять способ, когда публикаторов без учётной записи Git становится больше, чем команда успевает принимать запросами на изменение. То же — когда описание открыто, файл закрыт и заявки уже не разбираются почтой. API поиска нужен, когда наборов столько, что общий JSON тяжелеет для каждой программы, или когда сборщик по правилам страны принимает только API. Для наборов высокой ценности ЕС регламент требует и API, и массовую выгрузку: файл описаний закрывает выгрузку метаданных, API данных он не заменяет. Подробнее — в главе об интерфейсах доступа.

Сборка, которая не укладывается в предел хостинга, — тот же порог. На GitHub Pages это сайт больше 1 ГБ или сборка дольше 10 минут. Дальше берут портал или выносят файлы в хранилище, а на сайте оставляют описания.

Пример: каталог, который ведут двое

Контекст. Городская библиотека публикует около 80 открытых наборов: списки учреждений, расписания, годовые таблицы. Файлы остаются на сайтах ведомств. Региональный портал чужие организации не принимает. Правят двое сотрудников, у обоих есть учётная запись GitHub. Закрытых файлов нет. Описание должно находиться в поисковике. Если позже появится национальный сборщик, ему отдают один файл.

Варианты. Свой CKAN. Облачный PortalJS с ролями. JKAN. Обычный генератор и файл `data.json`, собранный скриптом.

Критерии. У каждого набора своя страница, адрес которой не меняется от правки заголовка. В файле описаний есть издатель, лицензия, дата и ссылка на файл. На странице есть schema.org. Сопровождение обходится без сервера приложений и без годовой платы. Оба редактора справляются сами.

Решение. JKAN, хостинг файлов — GitHub Pages. До первой публикации в шаблон добавляют schema.org, издателя и дату изменения в `data.json`. Форму Decap не включают: двоим проще править файлы, чем держать отдельный сервер входа. Если сборщик, когда он появится, ждёт DCAT-AP, шаблон переделывают под этот профиль заранее.

Почему. CKAN добавляет сервер при двух редакторах. Облачный PortalJS на первом платном тарифе даёт роли, которых здесь нет, и стоит от 1188 долларов в год; API поиска у него начинается со следующего тарифа. Обычный генератор тоже соберёт страницы, но страницу набора и `data.json` придётся писать с нуля. У JKAN они есть. Шаблон при этом не готов к публикации, пока в файле нет издателя: это видно по opendata.scot.

Последствия. Поиск на сайте фильтрует JSON в браузере. Программе отдают весь `data.json`. Файлы данных библиотека не копирует: пропажа файла на сайте ведомства видна как битая ссылка. Человек без учётной записи GitHub набор не добавит.

Когда пересмотреть. Появился третий публикатор, который не будет работать с GitHub. Понадобилось скрывать файл и оставлять описание открытым, и заявок стало больше, чем разбирает почта. Региональный портал начал принимать организацию библиотеки: два каталога сверяют и оставляют один, если дубли мешают. Сборка упёрлась в 1 ГБ или в 10 минут на GitHub Pages.

Как записать такую карточку у себя — в главе «Как обосновать решение».

Как проверить

Команды ниже повторяют сверку 14 октября 2026 года. Ожидаемые ответы сняты с живого opendata.scot. После правки шаблона список полей будет другим.

```
# Файл описаний открывается без входа. В тот день: 2596 наборов,
# схема DCAT-US 1.1, у первой записи нет ключа publisher
curl -s "https://opendata.scot/data.json" | python3 -c "import json,sys;
d=json.load(sys.stdin); print(len(d['dataset']), d['conformsTo']); print(sorted(d['dataset']
[0]))"

# На странице есть RDFa DCAT и нет schema.org
curl -s "https://opendata.scot/datasets/aberdeen+city+council-academy+school+catchments/" |
python3 -c "import sys; t=sys.stdin.read(); print('schema.org', 'schema.org' in t);
print('dcat', 'dcat:Dataset' in t)"
```

Первая команда печатает `2596`, адрес схемы `https://project-open-data.cio.gov/v1.1/schema` и список из шести ключей без `publisher`. Вторая печатает `schema.org False` и `dcat True`.

Дальше открывают свой сайт. В `data.json` у записи есть издатель, лицензия, дата изменения и прямая ссылка. Строка `conformsTo` совпадает с профилем, который ждёт сборщик: её сверяют с полями записи. В HTML страницы есть `schema.org/Dataset`. Адрес `sitemap.xml` отвечает. Файл набора, если он лежит у вас, открывается по прямой ссылке без входа.

Число наборов в национальном каталоге Чехии в тот же день считали так. Ответ был `31272`.

```
curl -sG "https://data.gov.cz/sparql" \
  --data-urlencode "query=SELECT (COUNT(DISTINCT ?d) AS ?n) WHERE { ?d a
<http://www.w3.org/ns/dcat#Dataset> }" \
  -H "Accept: application/sparql-results+json"
```

Источники

- JKAN, репозиторий, архитектура, установка и шаблон `data.json` на ветке `main`. Выпуск 2.0.0 — 30 марта 2023 года. Карта сайта — коммит `b4733fb` от 29 июля 2026 года, в тег 2.0.0 не входит. Последний push ветки — 7 сентября 2026 года. Коммиты по годам: пауза 2021 года, замена редактора на Decap в конце 2022 года. Сверено 14 октября 2026 года
- Схема набора DCAT-US 1.1 — обязательные поля, включая издателя, дату изменения, идентификатор и уровень доступа. Адрес схемы, который пишет JKAN, перенаправляет сюда
- opendata.scot и репозиторий — 2596 наборов, 5 878 536 байт в `data.json`, 5 775 288 байт в `datasets.json`, шесть полей, издатель пуст, 462 значения `No licence`, 232 набора без ссылки на файл, страница без schema.org. Синхронизация примерно раз в неделю, последний прогон 9 октября 2026 года. Проверено 14 октября 2026 года
- Репозиторий сайта data.gov.cz — Jekyll и GitHub Pages в README; каталог наборов вынесен в LinkedPipes DCAT-AP Viewer. Точка SPARQL — 31 272 набора, 133 каталога, 39 624 адреса скачивания, из них 16 на data.gov.cz. Запрос 14 октября 2026 года
- И. Бегтин, Telegram: портал открытых данных Чехии, 24.02.2023 — около 142 тысяч наборов, из них около 137 тысяч у кадастровой службы; opendata.scot на JKAN, 16.07.2022; Open Data Scotland, собранный сообществом, 13.05.2023
- PortalJS: бэкенды, публикация и тарифы — статический экспорт, сборка страниц из CKAN, тарифы 1188, 3588 и 14 990 долларов в год. Сверено 14 октября 2026 года. Репозиторий datopian/portaljs, лицензия MIT
- Open SDG, выпуск 2.4.0 от 19 августа 2025 года, лицензия MIT. Источники данных и SDMX, вывод SDMX в конфигурации. sdgdata.gov.uk — на странице есть `jekyll-open-sdg-plugins`, проверено 14 октября 2026 года
- Рекомендации STAC о статическом каталоге — файлы, относительные ссылки, каталог не отвечает на запросы. STAC Browser. Каталог Earth Engine: catalog.json, 133 дочерние ссылки, проверено 14 октября 2026 года
- Datasette, репозиторий simonw/datasette, лицензия Apache-2.0. Datasette Lite — база открывается в браузере, последний push 9 августа 2026 года
- Пределы GitHub Pages — 1 ГБ на опубликованный сайт, 10 минут на сборку, рекомендуемый 1 ГБ на репозиторий-источник, мягкий предел трафика 100 ГБ в месяц. Сверено 14 октября 2026 года
- Числа установок JKAN, OpenSDG и Datasette — выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года.

Git-репозитории

Репозиторий нужен, когда небольшую таблицу правят несколько человек и важно, кто что изменил. Историю Git показывает сам. Публикацией набор становится, когда рядом лежат описание колонок, лицензия и страница с постоянным адресом. Если на таблицу будут ссылаться в тексте, фиксируют выпуск и получают для него DOI. Файл больше предела площадки в сам Git не кладут: его прикладывают к выпуску или выносят в хранилище, а в репозитории оставляют адрес.

Площадки — GitHub, GitLab и другие. Цифры ниже сняты с документации GitHub.com и GitLab.com 14 октября 2026 года. На своём сервере пределы задаёт администратор, и перед публикацией большого файла их сверяют заново.

Когда репозитория хватает

Ситуация	Куда
Таблица до нескольких десятков мегабайт, правят двое-трое, нужна история строк	репозиторий и страница из тех же файлов
На годовой срез будут ссылаться в отчёте или статье	тот же репозиторий плюс выпуск с DOI
Файл больше предела Git	облако или вложение к выпуску, описание остаётся в репозитории. См. главу об облаке и хабах
Публикаторов много, и они не работают с Git	портал или уже существующий каталог. См. «Когда портал не нужен»
Нужен только DOI, историю правок никто не смотрит	общий репозиторий вроде Zenodo, без своего Git

Страницу, разметку schema.org и запись в чужом каталоге эта глава не собирает по шагам. Порядок — в главе «Как сделать публикацию находимой».

Что положить в репозиторий

Готовые файлы лежат в шаблонах. Их копируют и подставляют свои названия.

Файл	Зачем
`README.md`	что это, за какой период, как обновляется, кто отвечает
`LICENSE`	лицензия именно на данные
`datapackage.json`	список файлов и смысл колонок
`data/`	сами файлы в открытых форматах: CSV, JSON, Parquet
`CITATION.cff`	как сослаться. На GitHub по нему появляется кнопка «Cite this repository»

Один набор — один пакет. Пакет — это папка с описателем и файлами. Папкой может быть весь репозиторий или каталог внутри него, если наборов несколько и у каждого свой описатель. Десять несвязанных таблиц в одном репозитории без описателя — та же простыня ссылок, только в Git.

Описатель называется `datapackage.json` и лежит в корне пакета. Так требует стандарт Data Package 2.0: страница стандарта обновлена 25 августа 2026 года. Другое имя допустимо внутри своей системы, публиковать пакет под другим именем стандарт не советует. В шаблоне книги файл уже так назван, и в нём есть строка `$schema` со значением `https://datapackage.org/profiles/2.0/datapackage.json`. Без этой строки описатель проверяют по профилю 1.0. Поле `created` в профиле 2.0 — дата и время по RFC 3339, а не одна календарная дата. В шаблоне так и записано.

Живой репозиторий Country Codes этому правилу не следует: 14 октября 2026 года описатель лежал в корне как `datapackage.yml`. Колонки в нём описаны, файл CSV на месте, но опубликованное имя — не то, которое стандарт просит у пакета. На него можно смотреть как на образец словаря полей и как на напоминание сверить имя файла.

Если рядом с таблицей лежит скрипт, у кода и у данных часто разные лицензии. Их пишут явно: данные — в `LICENSE` и в `datapackage.json`, код — в шапке скрипта или во втором файле, на который README ссылается. Одна фраза «лицензия в корне» оставляет читателя гадать, на что она распространяется.

Скрипт, который собирает таблицу, тоже кладут в репозиторий. Обновление можно запускать автоматически: на GitHub это Actions, на GitLab — CI. Пропущенный квартал тогда виден по упавшему запуску.

Площадки и пределы

Git хранит текст и небольшие таблицы. Большой файл раздувает историю: его копируют при каждом клонировании, даже если он уже не нужен. Отсюда пределы.

	GitHub.com, бесплатный тариф	**GitLab.com, бесплатный тариф**
Файл в обычном Git	предупреждение, если файл больше 50 МиБ; отказ, если больше 100 МиБ. Через сайт — не больше 25 МиБ. Отдельно советуют держать один объект в пределах 1 МБ: таблица в несколько мегабайт пройдёт, клонирование и просмотр правки станут тяжелее	новый файл размером 100 МиБ и больше не принимают. Через сайт — не больше 100 МиБ
Весь репозиторий	лучше меньше 1 ГБ, настоятельно советуют меньше 5 ГБ. Размер папки `.git` рекомендуют держать в пределах 10 ГБ. Один push — не больше 2 ГБ	10 ГиБ на проект, вместе с Git LFS. Один push — не больше 5 ГиБ. На платном тарифе предел проекта — 500 ГиБ
Git LFS	файл до 2 ГБ. В месяц входят 10 ГиБ хранения и 10 ГиБ скачиваний на аккаунт, а не на каждый репозиторий. Квота кончилась, а бюджет на перерасход нулевой — пользование LFS блокируют до следующего месяца	место LFS входит в те же 10 ГиБ проекта. Сверх них проект становится только для чтения, пока размер не вернут в предел или не купят место
Вложение к выпуску	общий объём и трафик не ограничивают. Каждый файл — не больше предела LFS на тарифе, на бесплатном это 2 ГБ	в таблице настроек GitLab.com отдельный предел вложений к выпуску не назван. Файл, который не входит в 10 ГиБ проекта, в этот проект не кладут

МиБ и ГиБ — единицы из документации площадок: 1 МиБ = 1024 × 1024 байт. Страница квот GitLab.com пишет 10 ГиБ, таблица настроек того же сайта — «10 GB» для того же предела, размера репозитория вместе с LFS.

Git LFS — дополнение к Git. В истории остаётся короткая ссылка, сам файл лежит отдельно. Его ставят, когда файл должен приезжать вместе с клонированием и не влезает в обычный Git. Две оговорки. На GitHub Pages файлы LFS не отдают: страница, собранная из того же репозитория, большой файл не покажет. Квота скачиваний на бесплатном GitHub — 10 ГиБ в месяц на весь аккаунт. Популярный набор может исчерпать её, и чужой человек файл не получит.

Так устроен geoBoundaries: границы административных единиц лежат в Git LFS, это видно по файлу `.gitattributes` (папки `releaseData` и `sourceData`). Сайт и API — отдельные репозитории того же проекта. 14 октября 2026 года файл `LICENSE` называл лицензию CC BY 4.0, README рядом писал «CC BY 4.0 / ODbL», а карточка GitHub помечала лицензию как Other. Текст читают в файле, значок на карточке его не заменяет.

Для файла, который скачивают целиком раз в квартал, проще вложение к выпуску: на GitHub у него нет месячной квоты трафика. В репозитории тогда остаются README, описатель и адрес вложения. Ещё крупнее — облако, см. главу об облаке и хабах.

Площадка может сменить правила или стать неудобной в вашей стране. Копия выпуска в Zenodo эту зависимость уменьшает. На GitHub связка делает копию сама. На GitLab.com такой кнопки нет: архив выпуска загружают в Zenodo вручную или скриптом.

Почему репозиторий почти не ищут

У репозитория нет признака «здесь данные». Для площадки он такой же, как репозиторий с кодом. В 2022 году поиск по имени файла, выполненный из учётной записи, находил на GitHub десятки миллионов CSV, JSON, XML и баз SQLite. Каталогом он от этого не становится.

14 октября 2026 года поиск по коду без токена ответил `401` и текстом «Requires authentication». Поиск репозиториев по темам при этом открыт. Тема — метка, которую автор ставит сам. В тот день поиск насчитал 19 595 репозиториев с темой `dataset` и 10 590 с темой `open-data` (поле `incomplete_results` было `false`). Это только те репозитории, где автор метку поставил.

Человек и программа отличают репозиторий с данными по косвенным признакам: текст `README.md`, файлы в корне, `datapackage.json`. Темы `dataset` и `open-data` всё равно стоит поставить: это бесплатно и помогает тем, кто ищет по темам. Страница с разметкой schema.org нужна отдельно. Без неё поисковик данных репозиторий почти не видит.

Выпуск и DOI

Ссылка на файл в ветке `main` открывает сегодняшнее состояние. Завтра в той же ссылке будут другие строки. Выпуск фиксирует это состояние. На практике это метка версии, например `v2026.1`, и приложенный к ней архив. В отчёте ссылаются на выпуск, а не на ветку. Как устроены адрес и DOI, — в главе об идентификаторах.

На GitHub репозиторий связывают с Zenodo в профиле Zenodo. Связка забирает выпуски, созданные после включения, и каждому выдаёт DOI. У ряда выпусков есть и общий DOI: он открывает последнюю версию. В тексте цитируют DOI конкретного выпуска. Пока связки нет, из шаблона `CITATION.cff` убирают блок `identifiers` с пустым DOI, чтобы не цитировать

несуществующий номер. DOI, который Zenodo выдал, вписывают в `CITATION.cff` уже на ветке — для кнопки на GitHub. Переписывать ради этого сам архив выпуска не нужно: на странице Zenodo номер и так виден, а новый выпуск стал бы новой версией.

В `CITATION.cff` для данных нужна строка `type: dataset`. Без неё GitHub оформит ссылку как ссылку на программу. Кнопка при этом читает именно `CITATION.cff`.

Запись Zenodo устроена отдельно. Документация связки лежит в разделе про программы, а поддерживаемый пример `CITATION.cff` помечен `type: software`. Если в корне есть `.zenodo.json`, Zenodo читает только его и `CITATION.cff` для записи игнорирует. Справка с этим правилом опубликована 30 января 2026 года. Для набора данных в `.zenodo.json` ставят `upload_type` со значением `dataset` и те же название, авторов и лицензию, что в `CITATION.cff`. Код лицензии пишут так, как его знает Zenodo: `cc-by-4.0`, а не строка SPDX `CC-BY-4.0` из файла цитирования. Совпадает смысл лицензии, а не написание. Расхождение CC BY и CC BY-NC уже ломало публикацию без портала, пример — в главе «Когда портал не нужен».

```
{
  "title": "Название набора данных",
  "upload_type": "dataset",
  "license": "cc-by-4.0",
  "version": "2026.1",
  "creators": [
    {"name": "Фамилия, Имя", "affiliation": "Организация"}
  ]
}
```

После первого выпуска запись открывают и проверяют три вещи: тип — набор данных, лицензия совпадает с файлом `LICENSE`, DOI ведёт на этот выпуск. DOI нужен, если на данные ссылаются в тексте. Для внутренней таблицы его не заводят.

На GitLab автоматической связки нет. Метку и архив делают там же, файл загружают в Zenodo отдельным шагом и сверяют ту же тройку.

Форматы под тех, кто будет пользоваться

В репозитории легко выложить несколько форматов одного и того же списка. От этого зависит, возьмут ли данные. Два репозитория со списком стран показывают разницу. Country Codes отдаёт один CSV и словарь колонок. Country List отдаёт тот же по смыслу список на многих языках и сразу в файлах, которые программа подключает без преобразования. 14 октября 2026 года в каталоге `data/ru` лежали одиннадцать файлов: CSV, JSON, YAML, XML, HTML, простой текст, XLIFF, PHP и три варианта SQL — MySQL, PostgreSQL и SQLite. README перечисляет те же форматы. У первого репозитория в тот день было 981 звёзд, у второго — 5 253.

Звёзды — не оценка лицензии. У Country List файл `LICENSE` — текст MIT, он говорит о программе. У Country Codes в описателе указано общественное достояние ODC-PDDL-1.0, а README просит проверить ограничения ISO на коды стран, из которых список собран. Человек, которому нужен файл для базы, смотрит на формат. Человек, который вставляет список в продукт, читает лицензию в файле.

Вывод прежний. Словарь колонок нужен, и его дополняют форматами, в которых данными пользуются. Сначала выясняют, кто эти люди. Аналитик ждёт CSV или Parquet, разработчик — JSON или SQL. Как выбрать формат и почему GitHub показывает таблицей только CSV и TSV не больше 512 КБ, — в главе о форматах.

Код сборки вместо данных

Иногда публикуют не таблицу, а программу, которая её собирает. Так делают, когда источник выкладывает Excel, меняет файлы без предупреждения или не даёт прямой ссылки. Готовая таблица устареет при следующем обновлении, скрипт можно запустить снова.

В 2022 году так опубликовали сборщики реестров аккредитованных ИТ-компаний и отечественного ПО с сайта Минцифры России. 14 октября 2026 года в репозитории по-прежнему лежат описания сборки, а не сами наборы: в папке `itregistry` один файл `datacrafter.yml`, он указывает, откуда брать Excel и как назвать колонки. README репозитория — одна строка «Datacrafter examples».

Скрипт живёт, пока жив источник. Если ведомство уберёт файл или сменит адрес, код перестанет собирать таблицу, а копии у вас не останется. Поэтому рядом с кодом публикуют результат: выпуск с готовыми файлами или копию в Zenodo. Как сохранять данные, которые могут исчезнуть у источника, — в главе об архивации.

Совместная правка

Правку принимают запросом на слияние. На GitHub он называется pull request, на GitLab — merge request: это предложение изменить файлы, которое второй человек смотрит до того, как оно попадёт в основную ветку. В открытом репозитории на бесплатном GitHub можно сделать одобрение обязательным и назначить проверяющих файлом `CODEOWNERS`, например на папку `data/`. На GitLab.com в бесплатном тарифе одобрение слияние не блокирует. Как устроены роли и чем тарифы отличаются, — в главе о доступе. История коммитов при этом и есть журнал изменений.

Большую таблицу в окне запроса не прочитать. GitHub показывает разницу по одному файлу до 500 КБ или 20 000 строк, весь запрос — до 1 МБ или 20 000 строк. Проверку тогда делает программа, которая сверяет CSV со словарём колонок:

```
frictionless validate datapackage.json
```

Команду ставят в автоматический прогон при каждом изменении. На GitHub это Actions. Что именно ловит такая проверка, — в главе о полноте метаданных.

Версии таблицы, а не файла

Git сравнивает строки текста. Для таблицы на сотни тысяч строк этого мало: проверяющий хочет видеть, какие ячейки поменялись, а не два файла по 50 МБ.

Dolt — база данных с языком SQL, совместимая с MySQL, которую клонируют, ветвят, сливают и отправляют так же, как репозиторий Git. Публичные базы бесплатно хранит **DoltHub**. Правка проходит запросом на слияние, и разница показана по строкам и ячейкам. К истории можно обратиться обычным SQL: таблица `dolt_log` — журнал коммитов. DoltLab — та же площадка для установки у себя.

```
# Два последних коммита публичной базы, без входа
curl -s "https://www.dolthub.com/api/v1alpha1/dolthub/ip-to-country/master?
q=SELECT%20commit_hash%2C%20committer%2C%20date%20FROM%20dolt_log%20LIMIT%202"
```

15 октября 2026 года ответ начинался с `"query_execution_status":"Success"`. Ветка у этой базы называется `master`. По адресу с `main` сервер ответил «branch not found».

Это ступень между репозиторием и порталом. История правок как в Git, но для таблицы, которая уже не помещается в CSV на пару мегабайт. Риск тот же, что у любой площадки одного оператора: в 2022 году так закрылся каталог QRI с похожей идеей. Поэтому заранее проверяют, что базу вместе с историей можно забрать к себе командой `dolt clone`. Другие новшества того же ряда — в главе о передовых каталогах.

Пример: квартальная таблица двоих

Контекст. Два сотрудника ведомства раз в квартал правят таблицу расходов. Файл — CSV около 2 МБ. Нужно видеть, кто поменял строку. Раз в год таблицу цитируют в отчёте. Национальный каталог эту организацию не собирает. Своего сервера и договора с DataCite нет.

Варианты. Свой портал. Файл только в Zenodo, без репозитория. Репозиторий с README и без словаря колонок. Репозиторий с `datapackage.json`, лицензией и страницей; годовой выпуск уходит в Zenodo.

Критерии. История правки видна без сервера. У колонок есть имена и типы. Годовой файл, на который ссылается отчёт, не меняется задним числом. Сопровождение укладывается в работу этих двух человек.

Решение. Репозиторий на GitHub или GitLab, файлы — по шаблонам книги. Страница собирается из того же репозитория. Раз в год ставят метку выпуска. На GitHub выпуск забирает Zenodo. На GitLab архив загружают в Zenodo сами. В обоих случаях у записи тип «набор данных».

Почему. Портал добавляет сервер, а публикаторов по-прежнему двое. Один Zenodo не показывает, кто поправил ячейку в марте. Репозиторий без `datapackage.json` оставляет таблицу без смысла полей. Ветка `main` не годится для сноски в отчёте: по той же ссылке в следующем квартале будут другие числа.

Последствия. DOI ведёт на копию выпуска в Zenodo. Ссылку «таблица на сегодня» дают на страницу, ссылку из отчёта — на DOI этого года. Файл больше предела в этот репозиторий не кладут. Лицензия в `LICENSE`, в описателе и в записи Zenodo одна и та же.

Когда пересмотреть. Файл упёрся в предупреждение площадки (на GitHub — с 50 МиБ) или в отказ. Разницу в запросе на слияние уже нельзя прочитать: тогда смотрят на Dolt. Появилась третья группа, которая не работает с Git. Национальный каталог начал требовать файл описаний: его кладут рядом, портал из-за этого ещё не обязателен. Годовой срез цитируют чаще, чем его выпускают.

Как проверить

Проверка занимает несколько минут и не требует входа на площадку. Ожидаемые ответы ниже получены 14 октября 2026 года.

```
# Файл открывается без входа: код 200, длина 133966 байт
curl -sI "https://raw.githubusercontent.com/datasets/country-codes/main/data/country-
codes.csv"

# Описатель на месте. У этого репозитория он в YAML, стандарт просит datapackage.json
curl -sI "https://raw.githubusercontent.com/datasets/country-codes/main/datapackage.yml"

# Поиск по содержимому файлов без токена закрыт
curl -s "https://api.github.com/search/code?q=filename:datapackage.json"

# Темы репозиториев видны без входа. Смотрят поле total_count
curl -s "https://api.github.com/search/repositories?q=topic:dataset&per_page=1"
```

Первые два ответа — код 200. Третий — код 401 и сообщение, что нужна аутентификация. В четвёртом в тот день `total_count` был 19 595. Поле `last_modified` в описателе Country Codes стоит `2023-09-25`, а последний push репозитория на тот день — 1 октября 2026 года. Дату в описателе и дату файла сверяют оба: они уже разошлись.

Дальше открывают свой репозиторий. В корне пакета есть `datapackage.json`, `LICENSE` и `CITATION.cff` со строкой `type: dataset` и без пустого DOI. Файл, на который будут ссылаться, открывается по адресу метки выпуска, а не только по адресу ветки. Если включён Zenodo, у записи тип набора данных и та же лицензия, что в `LICENSE`.

Источники

- Data Package 2.0 — имя `datapackage.json`, свойство `$schema`, страница обновлена 25 августа 2026 года. Профиль: datapackage.org/profiles/2.0/datapackage.json. Проверка: Frictionless, validating data
- О больших файлах на GitHub — 50 МиБ, 100 МиБ, совет про 1 ГБ и 5 ГБ, вложения к выпуску. Пределы репозитория — 10 ГБ для папки `.git`, push 2 ГБ, объект 1 МБ / 100 МБ, разница в запросе на слияние. Git LFS — размер файла по тарифам, LFS не работает с GitHub Pages. Квоты LFS — 10 ГиБ хранения и 10 ГиБ скачиваний на бесплатном тарифе. Сверено 14 октября 2026 года
- Квоты хранения GitLab.com — 10 ГиБ на проект в бесплатном тарифе, 500 ГиБ на платном, вместе с LFS. Настройки GitLab.com — в таблице тот же предел указан как 10 GB, push 5 ГиБ. Предел файла на бесплатном тарифе — 100 МиБ. Сверено 14 октября 2026 года
- CITATION.cff на GitHub — `type: dataset`. Спецификация Citation File Format

- Zenodo: включение репозитория GitHub, CITATION.cff, `.zenodo.json`. Справка: при наличии `.zenodo.json` файл `CITATION.cff` на запись не влияет, опубликована 30 января 2026 года
- Отображение CSV и TSV на GitHub — 512 КБ, запятая и табуляция
- Поиск кода на GitHub — нужен вход. 14 октября 2026 года запрос API без токена вернул 401. Темы: `dataset` — 19 595 репозиториев, `open-data` — 10 590, `incomplete_results: false`
- Country Codes — `datapackage.yml`, лицензия ODC-PDDL-1.0 в описателе, оговорка про ISO в README, 981 звезда, CSV 133 966 байт. Country List — одиннадцать форматов в `data/ru`, LICENSE — MIT, 5 253 звезды. Проверено 14 октября 2026 года. Сравнение форматов предложено в посте И. Бегтина Открытые данные как дата-продукт, ноябрь 2024
- geoBoundaries и репозиторий — Git LFS в `.gitattributes`, LICENSE — CC BY 4.0, README — «CC BY 4.0 / ODbL», карточка GitHub — Other. Проверено 14 октября 2026 года
- datacrafter-examples — 14 октября 2026 года в `itregistry` лежит `datacrafter.yml`, не набор
- И. Бегтин, Telegram: GitHub как источник данных, 24.06.2022; GitHub — крупнейший, но плохо структурированный каталог, 08.04.2024; чего не хватает GitHub, чтобы стать репозиторием данных, 11.07.2024; как отличить репозиторий с данными, 16.09.2025; код сборки вместо данных, 07.07.2022
- Шаблоны этой книги: README, datapackage.json, CITATION.cff
- DoltHub, What is Dolt? — SQL, совместимость с MySQL, клонирование, ветки и слияние; SQL API. Проверено 15 октября 2026 года

Облачные хранилища и общие хабы

Большой файл можно опубликовать без своего портала. Для этого разделяют две задачи: где лежат байты и где лежит страница, по которой файл находят. Открытый бакет решает первую. Общий хаб часто решает обе, но по своим правилам и до своего предела. Общая папка по ссылке не решает ни ту ни другую.

Свой портал нужен, когда публикаторов много или верхний каталог требует роли и сбор чужих описаний. Пока набор один и публикует одна группа, хватает страницы, бакета или хаба. В любом из этих случаев нужны постоянный адрес, лицензия и копия вне той площадки, на которую будут ссылаться. Если хаб — единственное место, где лежит набор, его закрытие равно удалению.

Три места, которые путают

Бакет — контейнер в объектном хранилище: Amazon S3, Azure Blob Storage, Google Cloud Storage или такая же система на своём сервере (MinIO, Ceph). У каждого файла свой ключ и адрес по HTTP. Место почти не ограничено. Каталога у бакета нет: список ключей — не карточки наборов.

Общий хаб — чужая площадка, куда автор кладёт свой набор: Zenodo, Figshare, Hugging Face, Kaggle, Source Cooperative. Карточка и поиск уже есть. Хаб данных значит другое: несколько каталогов одной организации под общей вывеской. ML-хаб — вид общего хаба, на нём рядом с наборами лежат модели. Zenodo принимает данные любой дисциплины.

Общая папка — Google Drive, Dropbox, «Яндекс Диск» и похожие диски. Ссылка живёт, пока живёт учётная запись того, кто её выдал. Лицензии и постоянного адреса страницы нет, поисковик данных папку обычно не видит. Так данные не публикуют.

	Открытый бакет	**Общий хаб**	**Общая папка**
Что получает читатель	файл по адресу	карточку и файл	файл, если ссылка ещё жива
Кто платит за место и скачивание	владелец бакета, если набор не в программе спонсорства	площадка, пока набор в пределах её правил	владелец учётной записи
DOI	нет, его выпускают отдельно	у Zenodo, Figshare, Kaggle и по кнопке у Hugging Face	нет
Скачивание без входа	да, если бакет открыт на чтение всем	обычно да; у Kaggle файл отдают после входа	часто только после входа
Если это место закроется	пропадут файлы	пропадут файлы и карточка	пропадёт вместе с учётной записью

Когда портал сам держит файлы в облаке, читатель этого не видит: он нажимает «Скачать» на странице портала. Здесь другое. Бакета или хаба хватает вместо портала, а страница описания всё равно нужна. Как устроено хранилище внутри портала — в главе о хранилище.

Открытый бакет

Бакет подходит для снимков, выгрузок и моделей, которые не влезают в Git и в универсальный репозиторий. Рядом с файлами кладут то, без чего папка остаётся свалкой:

- `README.md` — что это, за какой период, как обновляется, кто отвечает;
- файл лицензии на сами данные, тем же текстом, что и на странице;
- контрольные суммы, файлом `SHA256SUMS` или в `datapackage.json`;
- постоянный адрес по HTTPS. Временная подписанная ссылка для статьи не годится: у Humanitarian Data Exchange она живёт минуты, подробнее в главе о хранилище.

Большие файлы кладут в форматах, которые читают по частям, не скачивая целиком. Таблицу — в Parquet. Снимок — в Cloud Optimized GeoTIFF: программа забирает нужное окно по сети. Многомерный массив, климат или результат модели — в Zarr, набор кусков с тем же свойством. Для снимков в бакет кладут ещё статический каталог STAC: JSON-файлы, по которым программа находит нужный кадр. Отдельный сервер каталога для этого не нужен, см. главу о статических сайтах.

Ссылку из статьи и из разметки schema.org ведут на страницу, которую ведёте вы. В поле прямой ссылки — адрес файла в бакете. Как собрать страницу — в главе «Как сделать публикацию находимой».

Кто платит за скачивание

Пока бакет просто открыт на чтение, хранение и исходящий трафик оплачивает владелец. Популярный набор на сотни гигабайт может стоить заметных денег именно из-за скачиваний, а не из-за места. Это считают до объявления, а не после первого счёта.

Часть провайдеров забирает эту плату на себя, но по заявке и на срок.

Программа спонсорства открытых данных AWS оплачивает хранение в S3 и передачу данных для принятых наборов. Участие длится два года с даты начала, затем прекращается, если AWS его не продлит. Если участие заканчивают не из-за нарушения условий, на то чтобы забрать данные, дают не меньше 90 дней. Дальше плату за счёт несёте вы. Условия разрешают изменить или закрыть программу. Если в заявке не указана лицензия, данные считаются опубликованными под CC BY 4.0. Условия последний раз обновляли 11 сентября 2025 года. Принятый набор описывают файлом YAML в репозитории Registry of Open Data on AWS: это список, а не портал с карточками. Как устроен такой список — в главе о списках наборов. 14 октября 2026 года в репозитории было 1241 такое описание.

Google Cloud сам размещает короткую витрину публичных наборов и платит за их хранение. Свой набор туда не отправляют формой: документация просит написать на почту программы. Консоль и командная строка Google даже для чужого публичного набора требуют вход, а прямая ссылка API — нет. В BigQuery за хранение публичных наборов тоже платит Google, а запросы сверх первого терабайта в месяц — тот, кто их выполняет. Это удобный способ читать чужие данные внутри облака, а не опубликовать свои.

Режим **оплаты трафика читателем** (Requester Pays) перекладывает плату за скачивание на того, кто читает. Для открытых данных он ломает публикацию: без учётной записи облака файл не отдают. 14 октября 2026 года из 1241 описания реестра AWS хотя бы один бакет с

такой пометкой был у шести наборов, среди них снимки Landsat и NAIP, лидар USGS и часть бакетов Sentinel-2. Ещё у двух наборов, Common Crawl и поездок такси Нью-Йорка, в описании стоит требование учётной записи AWS. Поле в реестре и живой запрос сверяют оба: бакет `sentinel-s2-l1c` в файле помечен как Requester Pays, но файл `GENERAL_QUALITY.xml` в тот же день открылся без ключа. Бакет `usgs-landsat` анонимный запрос отклонил.

Цифровая модель рельефа Copernicus на AWS в тот же день открылась без входа. Файл `readme.html` в бакете `copernicus-dem-30m` ответил кодом 200, а кусок вспомогательного файла — кодом 206: сервер умеет отдавать диапазон байтов. Лицензия набора записана в файле реестра.

Общие хабы

Площадка уместна, когда нужная аудитория уже там, когда нужен DOI, а своего репозитория и договора с DataCite нет, или когда файл больше, чем разумно держать на своём диске. Пределы ниже — на 14 октября 2026 года. Научные репозитории с ценами Dryad и Harvard Dataverse разобраны отдельно, в главе о научных репозиториях. Карточки, Croissant и пустые описания наборов для моделей — в главе о ML-каталогах.

Площадка	Для какого набора	Предел	Что даёт	Чего не будет
Hugging Face	наборы и модели, читатели уже на хабе	открытое место бесплатного аккаунта — по возможности; закрытые наборы — 100 ГБ; один файл не больше 500 ГБ	карточка, загрузка из кода, Croissant у части наборов, DOI по кнопке	гарантии бесплатных терабайт; старые файлы пропадают, если историю сожмут
Kaggle	то же, если набор влезает в предел	200 ГБ на набор и столько же на все закрытые наборы аккаунта; в корне не больше 50 файлов	карточка, блокноты, Croissant, DOI	скачивания без входа
Zenodo	данные к статье и выпуск репозитория	50 ГБ и 100 файлов на запись; один раз до 200 ГБ по запросу	DOI у каждой версии	Croissant; приёма набора, разрезанного на записи ради обхода предела
Figshare	то же, когда Zenodo мал	бесплатно 20 ГБ на учётную запись, файл до 20 ГБ, не больше 500 записей	DOI	бесплатного места на сотни гигабайт
Source Cooperative	большие открытые массивы, чаще геоданные	в документации площадки — сотни терабайт; публиковать можно по заявке	страница из `README.md`, скачивание без входа	DOI, версий для цитирования, закрытых файлов

У Hugging Face за первыми гигабайтами открытого места просят платный тариф или грант на хранение: площадка прямо пишет, что бесплатное место дают по возможности и защищаются от злоупотреблений. Файл больше 200 ГБ она просит резать на части, жёсткий потолок одного файла — 500 ГБ. История репозитория при этом не архив: место освобождают, сжимая историю, и старые файлы пропадают. DOI имеет смысл включить до того, как на набор начнут ссылаться: после этого удалить запись просто так нельзя. Подробнее — в главе об идентификаторах.

Число записей на Zenodo не ограничено. Один большой набор нельзя разрезать на много записей, чтобы обойти 50 ГБ: правила считают это обходом предела. Поиск через API не отдаёт каталог целиком — не дальше 10 000-й записи одного запроса, полной выгрузки файлом нет. Как забрать описания — в главе об интерфейсах доступа.

Figshare+ берёт разовую плату за объём сверх бесплатных 20 ГБ: 1225 долларов за каждые 250 ГБ, один файл до 5 ТБ. У университетов на своём Figshare пределы другие, их называет администратор репозитория. Для лаборатории без такого договора 500 бесплатных записей — потолок витрины, а не замена университетского репозитория.

Source Cooperative как образец

Source Cooperative — площадка некоммерческой организации Radiant Earth. На неё стоит смотреть как на образец устройства. Набор здесь называется продуктом: папка в объектном хранилище, открытая на чтение всем, плюс `README.md` в корне, который сайт показывает как карточку. Что писать в карточке, решает автор. Площадка советует указать описание, лицензию отдельно для текста и для данных, примеры доступа, контакт и DOI, если он уже есть. Сама площадка DOI не выпускает.

Продукт может занимать сотни терабайт. В документации площадки RapidAI4EO указан как пример больше 100 ТБ, а Earth Genome — как публикация больше 60 ТБ обработанных снимков. На GitHub такое не положить. 14 октября 2026 года площадка была в бета-версии: скачивать может любой без регистрации, публиковать — по заявке. Код сайта открыт. Ведёт площадку Radiant Earth, деньги на развитие даёт Taylor Geospatial, место в облаке — вклад AWS и Azure. Новые данные в Azure больше не кладут. Закрыть файл или брать плату за скачивание пока нельзя: документация относит это к будущим возможностям. Продукт можно убрать из поиска сайта, но по прямой ссылке он остаётся открыт.

Версий для цитирования нет. В бакетах включено хранение предыдущих копий файла, но неактуальные копии удаляют через 90 дней. Так страхуются от случайной перезаписи. Выпуск, на который ссылаются в статье, эти копии не сохраняют.

Адрес файла постоянный и открывается без ключа. У архива data.gov, который выложила библиотечная лаборатория Гарвардской школы права, описание лежит по адресу `https://data.source.coop/harvard-lil/gov-data/README.md`: 14 октября 2026 года оно открылось без входа, тип ответа — текст Markdown. Сама коллекция и то, как она упакована, разобраны в главе об архивации.

Слабое место устройства в том, что у каждого продукта своя структура папок и свои примеры кода. Документацию читают каждый раз. У набора на сотни гигабайт её всё равно читают, и здесь она хотя бы лежит рядом с данными.

Интернет-архив хранит копию сайта и файла надолго, но это архив, а не каталог с поиском по полям. Большие файлы он сохраняет плохо. Как им пользоваться и почему снимка страниц недостаточно — в главе об архивации.

Когда что выбирать

Ситуация	Куда
Таблица меньше 100 МБ, важна история правок	Git-репозиторий, см. главу о Git
Данные к статье, до 50 ГБ, своего репозитория нет	Zenodo
Набор для моделей, читатели уже на хабе	Hugging Face; Kaggle, если набор не больше 200 ГБ
Сотни гигабайт и терабайты, файлы читают по частям	Source Cooperative или программа спонсорства облака; страница — на вашем сайте
Нужен DOI на весь большой массив, а заявку на хаб не приняли	Figshare+ или репозиторий организации, который такой объём берёт
Много публикаторов, роли, сбор чужих каталогов	свой портал или уже существующий типовой, см. «Когда портал не нужен»

Пример: массив снимков без своего портала

Контекст. Научная группа. Один набор производных снимков, около 400 ГБ, в Cloud Optimized GeoTIFF, обновление раз в год. Читатели открывают нужные окна из кода и не скачивают архив целиком. На данные будут ссылаться в статье. Своего сервера, библиотекаря и договора с DataCite нет. Национальный портал собирает описания ведомств и такой массив не разместит.

Варианты. Свой портал. Весь набор в Zenodo. Kaggle. Hugging Face на бесплатном аккаунте. Figshare+. Свой публичный бакет без спонсорства. Заявка в Source Cooperative или в программу спонсорства AWS, страница — на сайте группы.

Критерии. Файл открывается по частям без входа и без счёта читателю. У статьи есть DOI на то, что можно воспроизвести. Группа унесёт данные, если площадка закроется. Плата за трафик не становится сюрпризом. Сопровождение укладывается в работу одного человека.

Решение. Заявка в Source Cooperative: массив открытый, геоданные, читают по частям. Параллельно — страница на сайте группы с разметкой schema.org и прямой ссылкой на продукт. В Zenodo кладут документированный образец до 50 ГБ и в описании прямо пишут, что это образец, а полный массив лежит по адресу страницы. Образец получает DOI, на который ссылается статья.

Почему. В Zenodo 400 ГБ не входят: потолок разового увеличения — 200 ГБ, а резать один набор на записи ради обхода предела правила запрещают. Kaggle ограничен 200 ГБ. Бесплатное место Hugging Face для такого объёма не обещано, аудитория площадки — модели, этот снимок там не ищут. Figshare+ за объём 251–500 ГБ по тарифу октября 2026 года стоит 2450 долларов разово. Это запасной путь, если грант требует DOI на каждый файл. Свой бакет без спонсорства оставляет группу плательщиком каждого скачивания. Свой портал добавляет сервер и не добавляет ни формата файлов, ни читателей.

Когда пересмотреть. Заявку не приняли — считают трафик и подаются в программу AWS или платят Figshare+, если DOI нужен на весь массив. Появилась вторая группа публикаторов, которая не готова вести файлы в облаке. До срока спонсорства, если выбрали AWS, осталось меньше полугода, а продления нет. Читатели не могут открыть файл без учётной записи облака.

Чем это рискует

Площадка меняет тариф, предел или правила, а иногда закрывается. Это уже случалось.

- **QRI**, публичный каталог данных, в январе 2022 года объявил о закрытии. Сайт qri.cloud перестал принимать наборы, 1 апреля 2022 года его обещали отключить, пользователей попросили заранее забрать данные.
- **data.world** начинался как открытый облачный каталог. Потом компания сделала ставку на корпоративный каталог, в 2025 году её купила ServiceNow, а 13 июля 2026 года открытое сообщество закрыли. Наборы сообщества больше нельзя скачать, данные участников удаляют. Корпоративный каталог при этом продолжил работать.
- **Microsoft** 6 июня 2024 года закрыла Planetary Computer Hub — среду, где исследователи считали в блокнотах на инфраструктуре проекта. Каталог данных и программный интерфейс остались. В том же обсуждении Microsoft писала, что продолжает размещать больше 50 ПБ. Блокноты и файлы из личных папок хаба нужно было забрать самим: 4 сентября 2024 года оставшиеся пользовательские данные удалили.
- **Zenodo** не закрывался, но в 2024 году ограничил глубину одного поискового запроса. Для того, кто хочет забрать все описания сразу, это тоже смена правил.

Четыре привычки заводят до загрузки:

- лицензия написана в карточке и в файле рядом с данными, одним и тем же текстом, а не «как принято на сайте»;
- выпуск, на который ссылаются, имеет DOI или другую копию вне площадки;
- адрес страницы, которую вы считаете постоянной, ведёте вы сами, а хаб указан как место скачивания;
- до загрузки проверено, что свои наборы можно выгрузить целиком, а не по одному, и что чужой человек скачивает файл без входа.

Описание остаётся открытым и тогда, когда сам файл большой и лежит в облаке. Срок спонсорства, если место оплачивает провайдер, записывают в карточку решения: два года AWS кончаются в известный день, а не «когда-нибудь».

Как проверить

Проверка занимает несколько минут и не требует учётной записи облака. Ожидаемые ответы ниже получены 14 октября 2026 года.

```
# Открытый бакет: описание читается без ключа, ответ 200
curl -sI "https://copernicus-dem-30m.s3.eu-central-1.amazonaws.com/readme.html"

# Файл отдают по частям: ответ 206 и заголовок Content-Range
curl -sI -H "Range: bytes=0-100" \
  "https://copernicus-dem-30m.s3.eu-central-
1.amazonaws.com/Copernicus_DSM_COG_10_N00_00_E006_00_DEM/AUXFILES/Copernicus_DSM_10_N00_00_E0
06_00_ACM.kml"

# Оплата читателем: анонимный запрос отклонён
curl -s "https://usgs-landsat.s3.us-west-2.amazonaws.com/?list-type=2&max-keys=1"

# Пометка в реестре и живой доступ могут разойтись: файл открылся без ключа
```

```
curl -sI "https://sentinel-s2-l1c.s3.eu-central-1.amazonaws.com/GENERAL_QUALITY.xml"

# На хабе описание лежит рядом с данными и открывается без входа
curl -sI "https://data.source.coop/harvard-lil/gov-data/README.md"
```

В первом ответе смотрят код 200 и тип содержимого. Во втором — код 206: так устроены форматы, которые читают по частям. В третьем сервер отвечает отказом и прямо пишет, что анонимные запросы к бакету с оплатой читателем не принимаются. Четвёртый запрос в тот день вернул 200, хотя в файле реестра бакет помечен Requester Pays: поэтому верят запросу без ключа, а не только полю в чужом списке. Пятый ответ — текст Markdown длиной несколько килобайт.

Дальше открывают страницу набора в браузере, где вы не вошли в площадку. Если файл не скачался, для открытых данных это уже ответ, даже когда карточка красивая. Лицензию сверяют в трёх местах: файл рядом с данными, карточка хаба, страница, на которую ведёт ссылка из текста. Расхождение этих трёх уже ломало публикацию без портала, пример — в главе «Когда портал не нужен».

Источники

- Registry of Open Data on AWS и репозиторий описаний; 14 октября 2026 года в `datasets/` был 1241 файл YAML, дерево Git API не обрезало. Пометки Requester Pays и AccountRequired посчитаны по этим файлам в тот же день
- Программа спонсорства открытых данных AWS, условия (обновлены 11 сентября 2025 года, срок участия — два года и не меньше 90 дней на то, чтобы забрать данные) и памятка по включению — программа покрывает хранение в S3 и передачу данных
- Публичные наборы Cloud Storage и публичные наборы BigQuery — Google платит за хранение, свой набор туда отправляют письмом, а не формой
- Бакеты с оплатой читателем. Проверено 14 октября 2026 года: Copernicus DEM открывается без ключа и отдаёт диапазон байтов; usgs-landsat анонимный запрос отклоняет; sentinel-s2-l1c в реестре помечен Requester Pays, файл при этом открылся
- Лимиты хранения Hugging Face — бесплатное открытое место по возможности, закрытые наборы 100 ГБ, файл не больше 500 ГБ; документация наборов Kaggle — 200 ГБ на набор
- Пределы размера Zenodo, справка опубликована 11 апреля 2026 года, сверена 14 октября 2026 года
- Figshare: пределы бесплатной учётной записи и Figshare+ — 1225 долларов за 250 ГБ, сверено 14 октября 2026 года
- Source Cooperative: документация — бета-версия, заявка, управление Radiant Earth; основные понятия — продукт, `README.md`, RapidAI4EO больше 100 ТБ; инфраструктура — открытые бакеты и удаление прежних копий через 90 дней. Проверено 14 октября 2026 года. Архив data.gov: страница harvard-lil/gov-data и файл README.md
- QRI, Winding Qri Down, январь 2022 года (копия в веб-архиве): приём наборов остановлен, сайт обещали отключить 1 апреля 2022 года
- data.world, A fond farewell — открытое сообщество закрыто 13 июля 2026 года, сверено 14 октября 2026 года

- Microsoft, Retiring the Hub on June 6th, 2024 — хаб закрыт 6 июня 2024 года, каталог данных остался, оставшиеся файлы пользователей удалены 4 сентября 2024 года
- И. Бегтин, Telegram: Source Cooperative и дата-продукты, 25.08.2026; закрытие QRI, 19.01.2022; data.world как каталог после входа, 18.07.2025; закрытие Planetary Computer Hub, 28.05.2024

Как сделать публикацию находимой

Набор находят по странице с постоянным адресом. Человек читает на ней описание и лицензию, поиск по данным читает разметку, сборщик каталога забирает отдельный файл, а ссылка из отчёта ведёт на выпуск с DOI. Репозиторий, запись в Zenodo и карточка в чужом каталоге называют тот же набор и ведут на эту страницу.

Ниже порядок для одного набора среднего размера. Как устроены сами площадки — в главах о статическом сайте, о Git и об облаке и хабах. Как то же самое платформа делает для тысяч страниц — в главе о находимости каталога.

Кто ищет и что он читает

Одного файла в открытом доступе мало: у каждого читателя свой вход.

Кто ищет	Что открывает	Что для него готовят
Человек со ссылкой	страницу	описание, лицензия, файл, дата, контакт
Обычный поисковик	HTML и ссылки с других сайтов	текст на странице и карта сайта
Поиск по данным, например Google Dataset Search	разметку на странице	schema.org типа Dataset, которую можно прогнать тестом
Сборщик национального или ведомственного каталога	файл в профиле, о котором вы договорились	`data.json` или DCAT-AP, с заполненными полями этого профиля
Автор отчёта или статьи	выпуск, который не меняется задним числом	метку версии и DOI этого выпуска

Темы репозитория `dataset` и `open-data` помогают тем, кто ищет репозитории на самой площадке. Поиск по данным их не читает. Почему так устроен GitHub — в главе о репозиториях.

Один главный адрес

Когда мест несколько, у набора всё равно один главный адрес: страница, которую ведёте вы. Остальные адреса на неё опираются.

Поле разметки	Куда смотрит
`url`	главная страница
`identifier`	DOI выпуска, после того как номер выдан
`sameAs`	другая страница того же набора, например запись Zenodo
`includedInDataCatalog`	каталог, в котором запись уже появилась

Google просит ставить полную разметку на главную страницу. Если разметка стоит и на копии, у копии в `sameAs` лежит адрес главной. Одно и то же значение `sameAs` у двух разных наборов склеивает их в один. Адрес поиска, ветки `main` и списка файлов в `sameAs` не ставят: это не страница набора.

Адрес страницы не собирают из заголовка. Заголовок правят, путь остаётся: `…/bus-stops`. Старые ссылки из статей и из индекса поисковика помнят путь, а не сегодняшнее название.

Поле `includedInDataCatalog` заполняют после того, как сборщик опубликовал запись. Имя каталога в разметке до этого обещает карточку, которой нет.

Что поставить на страницу

Разметка — блок `application/ld+json` в HTML страницы. Готовый текст лежит в schema-dataset.jsonld. Его вставляют в шаблон страницы. В README на github.com тот же текст остаётся абзацем для человека: тега `script` там нет, и поиск по данным его не читает.

Документация Google, сверенная 14 октября 2026 года, требует у типа Dataset два поля: `name` и `description`. Описание — от 50 до 5000 знаков. Более длинный текст поиск обрезает на 5000. Одно короткое предложение часто не дотягивает до 50. В описании пишут, что в таблице, за какой период и территорию, кто собрал и как часто файл меняется.

Остальные поля Google называет рекомендуемыми. Без них тест может пройти, а пользоваться набором нельзя. Лицензию, издателя и прямую ссылку на файл просят ещё и потому, что тип Dataset ставят и на мусор: об этом в главе о находимости каталога.

Поле	Что писать
`name`	название, по которому этот набор отличить от соседнего
`description`	от 50 до 5000 знаков, тот же смысл, что в тексте страницы
`license`	ссылка на конкретную версию, например CC BY 4.0. Ссылка на CC BY без номера версии для Google слабее
`distribution` с `contentUrl`	прямая ссылка на файл и его `encodingFormat`. Ссылка на раздел сайта файлом не является
`creator`, `publisher`	кто собрал и кто издаёт
`dateModified`	дата файла. Её меняют в тот день, когда меняется файл
`temporalCoverage`, `spatialCoverage`	период в виде `2020-01-01/2025-12-31` и территория. Если периода или территории нет, строки из шаблона удаляют
`identifier`	DOI. В шаблоне книги этого поля нет: его добавляют после выдачи номера
`sameAs`	вторая страница того же набора

Google пишет, что понимает на странице и schema.org, и DCAT. Проверка расширенных результатов принимает schema.org. Пример DCAT в той же документации помечен как неподдерживаемый тестом. Поэтому на своей странице ставят JSON-LD: его можно проверить до того, как давать ссылку людям. Атрибуты `dcat:Dataset` на странице JKAN этот тест не подтверждают. Файл `data.json` для сборщика — отдельный документ, его Google со страницы не читает.

Словарь колонок для людей и программ лежит в `datapackage.json`. Google отдельно называет экспериментальной разметку CSVW прямо на странице и предупреждает, что тест на такой блок ругается. Для таблицы хватает CSV по прямой ссылке и описателя пакета.

Страница отдаёт этот HTML сама, без входа и без капчи. Оболочка, которую дорисовывает скрипт в браузере, для робота пуста. Как это выглядит у больших сайтов — в главе о находимости каталога.

Карта сайта

Google просит карту сайта, чтобы находить адреса наборов, и `sameAs`, чтобы понимать, где копии. У одного набора в карте одна строка. У каталога из десятков страниц без карты робот может не дойти до каждой.

Файл `robots.txt` поисковик читает только из корня имени сайта. Так сказано в RFC 9309, и тот же предел разобран на геопорталах в главе о находимости каталога. У адреса `https://имя.github.io/bus-stops/` корень — `https://имя.github.io/robots.txt`. Файл `robots.txt` внутри каталога `bus-stops` этим правилом не является. Если корнем вы не распоряжаетесь, отдельный запрет из вашего каталога не появится. Карту тогда отправляют в Google Search Console по полному адресу `sitemap.xml`.

GitHub Pages сам разметку Dataset и карту сайта не пишет. Оба файла кладут в сборку сайта. Пределы хостинга и готовые генераторы — в главе о статических сайтах.

Файл для сборщика

Разметка одной страницы в национальный каталог не попадает. Сборщик забирает файл в том профиле, который принимает сейчас, а не в самой новой редакции схемы. data.gov по-прежнему собирает и DCAT-US 1.1, хотя версия 3.0 уже опубликована. Какой язык кому отдают — в главе о стандартах.

Файл готовят, когда сборщик есть и вы подали ему адрес. Регистрация сайта в Search Console этот адрес сборщику не сообщает. Как подать каталог на сбор — в главе о федерациях.

Файл, который объявляет схему и молчит об обязательных полях, сборщик может забрать таким, какой он есть. Короткий `data.json` у JKAN и записи без лицензии у небольшого ведомства разобраны в главах о статических сайтах и «Когда портал не нужен». Если сборщика нет, отдельный файл ради строки `conformsTo` не заводят: страница с разметкой уже описывает набор.

Вторая страница с DOI

DOI нужен, когда на выпуск ссылаются в тексте. Для внутренней таблицы его не заводят. Как связать репозиторий GitHub с Zenodo и почему `.zenodo.json` перекрывает `CITATION.cff` — в главе о репозиториях. Пока номера нет, блок `identifiers` из шаблона `CITATION.cff` удаляют.

Запись Zenodo с типом «набор данных» сама становится страницей с разметкой Dataset. DataCite пишет, что в Google Dataset Search попадают посадочные страницы с классом Dataset, а согласование содержимого по DOI — только если запись в состоянии Findable и тип ресурса Dataset. Сроком, когда Google дойдёт до страницы, DataCite не распоряжается. Тип «программа» этот класс не получает: в `.zenodo.json` для данных ставят `upload_type` со значением `dataset`.

Живая запись показывает, что страница Zenodo и ответ по DOI — два разных документа. 14 октября 2026 года выпуск Cantometrics, DOI 10.5281/zenodo.4898406, отдал в HTML один блок JSON-LD. Обычный запрос страницы вернул код 200. Тот же адрес с заголовком браузера `Mozilla/5.0` вернул 403: отказ относится к этому заголовку, а не к отсутствию страницы.

- Тип в разметке страницы — `https://schema.org/Dataset`. Короткого слова `Dataset` сразу после кавычки там нет.
- Видимый текст описания — «The data repository for the Cantometrics dataset», 48 знаков. Строка в разметке длиннее, 56 знаков, только из-за тегов абзаца. Порог Google — от 50 знаков самого описания.
- `license` ведёт на текст CC BY-NC 4.0. В репозитории той же метки файл лицензии — CC BY 4.0. Расхождение уже разобрано в главе «Когда портал не нужен».
- `url` — `https://zenodo.org/records/4898406`. Поля `sameAs` нет.
- `contentUrl` ведёт на ZIP выпуска. Запрос заголовков этого адреса вернул код 200 и тип `application/octet-stream`. В разметке у файла указан `application/zip`.

Запрос JSON-LD по тому же DOI в тот день вернул другой документ: тип Dataset, описание из 48 знаков, без ссылки на файл. Поле `url` в нём — старый путь `/record/4898406`, он отвечает кодом 301 и уводит на `/records/4898406`. Адрес метки в GitHub этот ответ называет статьёй, на которой набор основан, а не второй страницей набора.

Отсюда практический вывод. DOI и страница Zenodo закрывают ссылку из отчёта и дают поиску по данным ещё одну страницу. Текст описания, лицензию и прямую ссылку держат на своей странице и в запись Zenodo копируют теми же. В `identifier` своей разметки ставят DOI выпуска. В `sameAs` — адрес записи Zenodo. В цитате используют DOI выпуска, а общий DOI всех версий оставляют тем, кому нужна последняя.

Порядок на одном наборе

Пример — расписание автобусов района. Имена вымышленные, файлы — шаблоны этой книги. Рядом можно открыть живой репозиторий Country Codes: у него другая тема, описатель называется `datapackage.yml`, а 14 октября 2026 года в HTML страницы репозитория не было блока JSON-LD.

1. **Репозиторий.** Положите шаблоны: README, `datapackage.json`, `LICENSE`, таблицу в `data/`. В `CITATION.cff` оставьте `type: dataset` и удалите блок `identifiers`, пока DOI нет. В README напишите, что это за расписание, за какой сезон, кто отвечает и как часто файл меняется. Темы `dataset` и `open-data` поставьте: они помечают репозиторий на площадке. Что ещё лежит в пакете и где кончаются пределы Git — в главе о репозиториях.

2. **Страница.** Соберите сайт из этого репозитория или положите страницу на уже существующий сайт. Путь `…/bus-stops` не меняется вместе с заголовком. В HTML вставьте блок из шаблона разметки. Описание сделайте длиннее 50 знаков, например: «Расписание остановок и времени отправления автобусов района за сезон: остановка, маршрут и время. Файл обновляют раз в сезон.» Лицензия — ссылка на CC BY 4.0. В `contentUrl` — прямая ссылка на CSV. Поле `identifier` пока не добавляйте.

3. **Проверка разметки.** Адрес страницы вставьте в проверку расширенных результатов. Ошибки типа Dataset исправьте до того, как давать ссылку. Страница должна открываться без входа: иначе тест и поисковик видят заглушку.
4. **Карта сайта.** Положите `sitemap.xml` с адресом этой страницы. Если распоряжаетесь корнем имени сайта, укажите карту в `robots.txt` этого корня. Иначе отправьте полный адрес карты в Search Console.
5. **Версия и DOI.** Если на расписание будут ссылаться, поставьте метку, например `v2026.1`, и получите DOI выпуска. Номер впишите в `CITATION.cff` на ветке и в `identifier` разметки. В `sameAs` поставьте адрес записи Zenodo. Лицензия в `LICENSE`, в описателе, на странице и в записи Zenodo — одна и та же. Для таблицы, на которую никто не ссылается, шаг пропускают.
6. **Сборщик.** Если региональный или национальный каталог забирает файл, положите его в профиле, который каталог принимает, и подайте адрес. Поля, которые профиль считает обязательными, заполните до заявки.
7. **Через несколько дней и ещё раз через несколько недель.** Найдите название в Google Dataset Search. Google пишет, что обход занимает несколько дней, а исчезновение из выдачи — дни или недели. Раньше этого срока площадку не меняют. Сначала перепроверьте тест разметки и то, что страница открыта без входа.

Пример: главный адрес на своей странице

Контекст. Двое уже ведут квартальную таблицу в репозитории. Раз в год ставят выпуск и получают DOI. Национальный каталог организацию не собирает. Через месяц по названию таблицу не находят в поиске по данным. Один участник предлагает давать ссылку на репозиторий, другой — только DOI, третий — поставить портал.

Варианты. Вкладка репозитория на github.com. Только запись Zenodo. Своя страница с разметкой, DOI в `identifier` и адресом Zenodo в `sameAs`. Портал.

Критерии. Разметку можно проверить до индексации. Ссылка из отчёта открывает неизменный выпуск. Человек без DOI открывает описание и файл. Сопровождение остаётся у этих двоих.

Решение. Главный адрес — своя страница. DOI выпуска стоит в отчёте и в поле `identifier`. Запись Zenodo — вторая страница того же набора.

Почему. Вкладка github.com у репозитория со словарём колонок не содержит Dataset: 14 октября 2026 года так открылась страница Country Codes. Запись Zenodo тип Dataset ставит сама, но описание и лицензию берёт из того, что ей отправили. У Cantometrics видимый текст короче порога Google, а лицензия на странице Zenodo разошлась с файлом в репозитории. Портал добавляет сервер при одном наборе и двух публикаторах.

Последствия. Карту сайта и разметку обновляют вместе с файлом. Пустой DOI не публикуют. Поиск по данным может сначала показать страницу Zenodo: у неё тоже тип Dataset. После индексации своей страницы обе страницы связаны через `identifier` и `sameAs`.

Когда пересмотреть. Наборов стало больше, чем двое успевают размечать. Появился сборщик и назвал профиль файла. Описание или лицензия на своей странице и в Zenodo разошлись.

Как записать такую карточку у себя — в главе «Как обосновать решение». Выбор репозитория для этой таблицы — в главе о Git.

Как проверить

Проверка своей страницы занимает несколько минут. Подставьте её адрес.

```
PAGE=https://example.org/bus-stops

# Страница открывается без входа
curl -sI -o /dev/null -w "%{http_code}\n" -L "$PAGE"

# В HTML есть тип Dataset: и короткое имя, и полный адрес schema.org
curl -sL "$PAGE" | grep -o '"@type"[[:space:]]*:[[:space:]]*"[^"]*Dataset"' | head -1
```

Код 200 и строка с `Dataset` значат, что страница отдаёт разметку. Дальше адрес вставляют в проверку расширенных результатов: документация Google требует у описания от 50 до 5000 знаков, и тест сверяет разметку с этими правилами. Пустого `identifier` в блоке быть не должно.

Команды ниже сняты 14 октября 2026 года с чужих публикаций. Они показывают разрыв, а не образец.

```
# Страница репозитория открывается. Блока JSON-LD в ней нет: вывод пустой
curl -sL -o /dev/null -w "%{http_code}\n" "https://github.com/datasets/country-codes"
curl -sL "https://github.com/datasets/country-codes" | grep -o 'application/ld+json' | head

# JSON-LD по DOI: тип Dataset, старый путь /record/, без distribution
curl -sL -H "Accept: application/ld+json" "https://doi.org/10.5281/zenodo.4898406"

# Старый путь перенаправляет на /records/
curl -sI -o /dev/null -w "%{http_code}\n" "https://zenodo.org/record/4898406"

# Страница Zenodo: код и тип Dataset в HTML
curl -sL -o /dev/null -w "%{http_code}\n" "https://zenodo.org/records/4898406"
curl -sL "https://zenodo.org/records/4898406" | grep -o '"@type"[[:space:]]*:[[:space:]]*"
[^"]*Dataset"' | head -1

# API записи и файл из contentUrl
curl -sL "https://zenodo.org/api/records/4898406"
curl -sI -L -o /dev/null -w "%{http_code}\n" \
  "https://zenodo.org/api/records/4898406/files/theglobaljukebox/cantometrics-v0.1-
alpha.zip/content"
```

В тот день страница Country Codes вернула 200, поиск `application/ld+json` ничего не нашёл. Ответ по DOI был типом Dataset. Путь `/record/4898406` вернул 301. Страница `/records/4898406` вернула 200, а поиск типа нашёл `"@type": "https://schema.org/Dataset"`. API записи вернул 200 и тип `dataset`. Заголовки файла из `contentUrl` вернули 200 и тип `application/octet-stream`. Запрос этой же страницы с заголовком `Mozilla/5.0` вернул 403.

Где путь ломается

- **Вкладка репозитория вместо страницы.** Файл и README на месте, типа Dataset в HTML нет. Обычный поиск может найти README. Поиск по данным ищет разметку.
- **Описание короче 50 знаков.** Тип Dataset объявлен, обязательное поле по правилам Google не заполнено. Теги вокруг короткой фразы длину текста не заменяют.
- **Пустой или вымышленный DOI.** Номер-заглушка из `CITATION.cff` попадает в чужие ссылки. Пока номера нет, блока `identifiers` и поля `identifier` нет.
- **Один `sameAs` у двух наборов.** Поиск склеивает разные таблицы в одну запись.
- **`contentUrl` ведёт на раздел сайта.** Скачивать нечего. В поле стоит адрес файла.
- **Файл объявляет профиль и пропускает его поля.** Сборщик публикует дырявую карточку. Без сборщика такой файл не заводят.
- **Три места с тремя лицензиями.** Человек по DOI и человек по репозиторию получают разные условия. Лицензию пишут один раз и копируют.
- **Капча, вход или страница из скрипта.** Тест разметки и робот видят заглушку. Через месяц индекс будет пуст по той же причине.

Чем этот путь отличается от портала

Портал сам собирает выгрузку, карту сайта и часто разметку для каждой страницы. Здесь это делает автор страницы, по одному набору. Проверка та же, что в главе о находимости каталога: адрес открыт, разметка проходит тест, карта сайта известна поисковику, нужный профиль отдан сборщику. Когда наборов становится больше, чем один человек успевает описывать, или сбор идёт по расписанию из многих организаций, возвращайтесь к главе «Когда портал не нужен»: порог там записан заранее.

Источники

- Разметка Dataset в документации Google — обязательны `name` и `description` от 50 до 5000 знаков; schema.org и DCAT на странице; пример DCAT проверка расширенных результатов не принимает; `sameAs`, карта сайта, `contentUrl`. Сверено 14 октября 2026 года
- DataCite: посадочные страницы и Google Dataset Search — класс Dataset на странице, карта сайта или ссылки, состояние Findable и тип ресурса Dataset для согласования содержимого. Справка DataCite по-прежнему называет старый инструмент проверки структурированных данных. Актуальная проверка в документации Google — Rich Results Test
- RFC 9309 — `robots.txt` читается из корня имени сайта
- Google Search Console, Google Dataset Search
- Country Codes — 14 октября 2026 года страница ответила кодом 200, блока `application/ld+json` нет. Имя описателя `datapackage.yml` разобрано в главе о Git-репозиториях

- Cantometrics, DOI 10.5281/zenodo.4898406 — 14 октября 2026 года. В HTML страницы JSON-LD с типом `https://schema.org/Dataset`, видимое описание 48 знаков, лицензия CC BY-NC 4.0, `contentUrl` на ZIP, `sameAs` нет. Обычный запрос страницы — 200. Запрос с заголовком `Mozilla/5.0` — 403. API записи — 200, тип `dataset`. JSON-LD по DOI — без `distribution`, `url` с путём `/record/`, этот путь отвечает 301. Заголовки файла из `contentUrl` — 200, `application/octet-stream`
- Шаблоны: /templates/README.md, /templates/datapackage.json, /templates/CITATION.cff, /templates/schema-dataset.jsonld
- Экспорт метаданных, находимость каталога, GitHub и Zenodo

ЧАСТЬ IX.

Практики

Хорошие практики, антипаттерны, жизненный цикл каталогов данных и их архивация.

1. Хорошие практики
2. Антипаттерны
3. Жизненный цикл каталога
4. Архивация и спасение каталогов

Хорошие практики

Хорошая практика — та, без которой набор нельзя найти или нельзя использовать. Остальное зависит от типа каталога. Ниже шесть вещей для любого способа публикации, почему каждая обязательна и кто это уже делает. Сбои, в которые превращается пропуск, собраны в главе об антипаттернах.

Что сделать	Почему без этого набор не берут	Где это уже видно
Файл открывается без платной программы, кодировка и разделитель написаны	иначе каждый пользователь чинит таблицу заново и по-своему	открытые форматы и словарь полей, ниже
У колонок есть словарь	колонка `sum_2` формально открыта и закрыта по смыслу	schema.data.gouv.fr, Table Schema
В карточке стоит частота, и она про файл, а не про правку описания	на набор строят сервис и должны знать, когда ждать новые строки	HDX: срок обновления считает сам портал
Лицензия, издатель, человек для ответа, период и территория — на наборе	без них нельзя ни законно использовать файл, ни понять, подходит ли он	балл карточки на data.gouv.fr
Весь каталог забирается программой в том профиле, который уже читает верхний портал	поле «по-своему» для сборщика не существует	постраничный DCAT и еженедельные CSV data.gouv.fr
Адрес набора не меняется, и есть куда написать о поломке	ссылка из отчёта и чужая программа живут дольше платформы	переход data.gov в 2026 году; обсуждения на data.gouv.fr

В техническом задании эти шесть пунктов помечают как обязательные. Чек-лист для остальных развилок — в главе о требованиях. Полный перечень по типам — в карте возможностей.

Файл

Формат открывается без платной программы. CSV, JSON, XML, GeoPackage, Parquet. PDF оставляют человеческой копией рядом: таблицу из него перепечатывают или распознают, и каждый делает это с новыми ошибками. Три вопроса, по которым файл годен для программы, — в главе о форматах.

У таблицы есть словарь полей. Название колонки, единица измерения, допустимые значения, что означает пустая ячейка. Словарь кладут рядом отдельным файлом или описывают так, как принято в отрасли: Table Schema, переменные DDI, структура SDMX. Формы и то, как французский портал проверяет файл по общей схеме темы, разобраны в главе о полноте метаданных.

Кодировка, разделитель и десятичный знак написаны. В карточке файла или в самом словаре: UTF-8 или другая кодировка, запятая или точка с запятой, точка или запятая в числах. Иначе половина таблиц одного издателя открывается, половина — нет, и нигде не сказано, где какая. Так выглядел раздел открытых данных Росстата; разбор — среди антипаттернов.

Один набор — то, что ищут по имени. Год и район, которые можно сделать колонкой, не заводят отдельными карточками ради счётчика. Делить стоит то, что ищут само по себе: бюджет города, временной ряд с постоянным адресом. Проверочный вопрос: станет ли

удобнее, если склеить эти файлы в один. Если да, это дробление. Примеры и граница — в той же главе об антипаттернах.

Частота в карточке — про файл, и её кто-то сверяет. Дата обновления — дата строк, а не день, когда поправили описание. Пользователь строит на наборе расчёт и должен знать, когда ждать продолжение.

На Humanitarian Data Exchange частота — обязательное поле из короткого списка: каждый день, каждую неделю, раз в год, а также «в реальном времени», «по необходимости» и «никогда». Портал сам считает срок, к которому ждут следующий файл. С апреля 2025 года правило такое: к конечной дате периода прибавляют заявленную частоту. Набор профилей стран EM-DAT обновляют раз в неделю. 14 октября 2026 года период данных заканчивался этим же днём, а срок следующего обновления стоял 21 октября. Частоту «каждую неделю» в тот день имели 866 наборов: поле не пустует. Когда срок выходит, письмо уходит ответственному за набор, а не в общий ящик портала. Сами авторы HDX просят выбирать ближайший регулярный срок, а не «по необходимости»: иначе проверка не срабатывает.

Карточка

Лицензия, издатель, контакт, период и территория стоят на каждом наборе, не в подвале сайта.

- **Лицензия** — ссылка на текст условий. Без неё файл можно скачать и нельзя законно использовать. Какую ставить — в главе о лицензиях.
- **Издатель** — организация, которая отвечает за содержимое. По нему ищут все наборы одного ведомства и отличают копию от первоисточника.
- **Контакт** — человек или команда с адресом, которые ответят про ошибку в строках. Ящик `info@` без фамилии этого не закрывает: письмо некому прочитать.
- **Период и территория** — чтобы решить, подходит ли набор, не скачивая его. «Данные за все годы по всей стране» в описании не заменяют заполненные поля: программу по ним не отфильтровать.

Сверх этого списка нужен один профиль, который уже читает ваш верхний портал: DCAT-AP, ISO 19115, DataCite, DDI или SDMX. Не все сразу, а тот, который соответствует типу. Верхний портал и поисковик читают описание программой, и поле с местным именем для них пустое. Как выбрать язык описания — в главе о стандартах.

Копия указывает на оригинал. Город, регион и страна могут показывать одну карточку, но у копии есть ссылка на источник. Иначе ошибку правят в трёх местах, а поиск считает один набор тремя.

data.gouv.fr показывает издателю балл карточки, чтобы дыры было видно до жалобы пользователя. Руководство портала перечисляет, что в балл входит: достаточно длинное описание, описанные файлы, заполненная и соблюдённая частота, открытая лицензия, хотя бы один открытый формат, территория и уровень её детализации, период. Балл с 2022 года экспериментальный: веса могут изменить, проверку по схеме данных в него ещё не включили. 14 октября 2026 года карточка самого каталога портала проходила семь проверок из девяти, которые отдаёт API. Не были закрыты описание файлов и временной охват.

Лицензия при этом стояла — французская открытая Licence Ouverte 2.0, — частота была заполнена и соблюдена, территория указана. Балл смотрит на карточку, не на строки внутри файла. Пустой файл с заполненными полями здесь получит высокий балл; как отличить одно от другого — в главе об уровнях качества.

Каталог, который забирают

Без выгрузки всех описаний каталог не попадёт в федерацию. Без постоянного адреса на набор нельзя сослаться. Без регистрации о каталоге не узнают те, кто ищет порталы, а не отдельные наборы.

Выгрузка открывается без входа и без договора. Национальному порталу одним файлом уже тесно, и это нормально, если программа может обойти страницы. 14 октября 2026 года выгрузка DCAT data.gouv.fr содержала 74 616 наборов, по 100 на страницу, всего 747 страниц: в файле указаны общее число и адрес следующей страницы. Рядом портал публикует весь каталог таблицами. В карточке частота — еженедельно. Утром того же дня там лежали свежие CSV: список наборов на 276 МБ и список файлов на 1,1 ГБ, плюс организации, обсуждения, примеры использования и сборщики. Любой агрегатор забирает это без отдельной договорённости. Как устроены другие способы — один файл, протокол, лента изменений — в главе об API и выгрузке.

Адрес переживает смену платформы. Его проектируют до переезда: старый путь либо остаётся, либо перенаправляет на новый. data.gov при уходе с CKAN на собственную разработку в апреле 2026 года оставил прежний каталог по отдельному адресу на переходный срок, чтобы пользователи успели перенастроить программы. Как удержать адрес — в главе об идентификаторах.

API поиска включают, когда наборов сотни и больше. Для десятка хватает страниц и одного файла описаний. Описания при этом читаются без регистрации и капчи: ограничивать можно файл, не карточку.

Есть куда сообщить об ошибке, и письмо доходит до владельца данных. Форма на странице набора, адрес в карточке или открытый трекер. На data.gouv.fr у набора есть обсуждение, а примеры использования — приложения и статьи, которые читатели сами привязали к карточке, — портал отдаёт отдельным файлом в той же еженедельной выгрузке. Обратная связь тогда видна и людям, и тем, кто забирает каталог целиком. Как это устроено на французском портале в целом — в главе о порталах открытых данных.

Каталог записан там, где его ищут. Национальный и тематический реестр, а для научного репозитория — re3data. Новость на своём сайте этого не заменяет. Куда регистрироваться — в главе о находимости.

Что добавляет тип

Шести пунктов хватает странице ведомства и городскому порталу. Тип каталога добавляет своё. Это не второй обязательный список «на всякий случай»: лишнее для вашего типа замедляет публикацию и не нужно пользователю.

Тип	Что добавить к шести пунктам	Где разобрано
Портал открытых данных	профиль страны, чаще всего DCAT-AP, и общая схема для тем, которые публикуют многие	стандарты, федерация
Научный репозиторий	DOI, версии, готовая ссылка для статьи, выдача описаний по OAI-PMH	научные репозитории
Геопортал	слой можно скачать, или прямо сказано, что доступен только просмотр; описание в ISO 19115	геопорталы
Каталог показателей	методика рядом с рядом; SDMX или другая массовая выгрузка таблицы, а не только экран	каталоги показателей
Каталог микроданных	уровни доступа, словарь переменных, описание открыто даже когда файл закрыт	микроданные
ML-каталог	описание в Croissant, неизменное деление на обучающую и тестовую части, лицензия и происхождение данных	ML-каталоги
Публикация без портала	страница открывается без входа, лицензия на месте, разметка schema.org с описанием от 50 знаков, дата и контакт	как сделать находимым

Диаграмма на сайте эти пункты не заменяет. Скачать по-прежнему можно файл или выгрузку.

Набор как продукт

Практики выше можно выполнить формально: карточка заполнена, файл открывается, а пользоваться набором неудобно. Помогает другой взгляд — относиться к набору как к продукту, у которого есть пользователи. В компаниях это называют «данные как продукт». Чем дата-продукт отличается от набора и какие каталоги устроены под продукты, разобрано в главе о единицах публикации. Для открытых данных подход пока редкий. Проверка одного набора выглядит так.

1. Известно, кто пользуется набором и для какой задачи. Хотя бы один сценарий, даже если файл бесплатный.
2. Есть ответственный, который понимает и данные, и пользователей. Он отвечает на вопросы и правит набор по обратной связи. Не «отдел», а человек или команда с адресом.
3. Документация опубликована и совпадает с файлом: схема, пример кода, ограничения. На HDX для ограничений есть отдельное поле карточки.
4. Данных можно получить больше чем одним способом: файл, API, срез. Формат выбран под пользователя. Аналитику нужен CSV или Parquet, разработчику — API или JSON. Пример, как это меняет спрос на одинаковые по сути данные, — в главе о Git-репозиториях.
5. Качество описано и проверяется постоянно, а не один раз перед первой публикацией и не по письму пользователя.
6. Есть версии и график обновления, и график соблюдают. Для научного набора версия — то, на что ссылается статья. Для оперативного — хотя бы дата файла и прежние выпуски, если на старые уже могли сослаться.
7. Дашборд, если он есть, стоит рядом с данными.

Пункт 6 совпадает с частотой из этой главы. Остальное — то, чего не даёт даже аккуратно заполненная карточка. О том же говорит идея «третьей волны» открытых данных: публиковать то, что нужно для конкретных задач, — см. наборы высокой ценности.

Откуда взялись эти правила

Практики не выдуманы под один портал. Их повторяют три документа, на которые опираются национальные политики открытых данных и научные репозитории. В книге они разложены по типам: то, что обязательно научному архиву, лишнее оперативному порталу.

Практика из этой главы	Хартия открытых данных, 2015	Data on the Web Best Practices, W3C, 2017	Принципы FAIR, 2016
Открытый формат и словарь полей	доступность и пригодность к использованию; сопоставимость	форматы и словари	совместимость
Частота и полнота	своевременность и полнота	качество данных	—
Лицензия на наборе	открытость по умолчанию	лицензии	повторное использование
Выгрузка всего каталога	доступность	доступ и метаданные	находимость
Постоянный адрес	—	идентификаторы	находимость
Куда сообщить об ошибке	улучшение управления и вовлечение граждан	обратная связь	—

Международная хартия открытых данных — шесть принципов для государств: открытость по умолчанию, своевременность и полнота, доступность и пригодность к использованию, сопоставимость и совместимость, улучшение управления и вовлечение граждан, инклюзивное развитие и инновации. На сайте хартии 14 октября 2026 года было написано, что её приняли больше 29 национальных и 146 региональных и городских правительств. Хартия говорит, что публиковать, и почти не говорит, каким файлом.

Data on the Web Best Practices — рекомендация W3C от 31 января 2017 года, 35 практик: от описания и лицензии до версий, сохранения и обратной связи. Это самый близкий к этой главе перечень. Он про данные в вебе вообще, не только про государственные порталы.

Принципы FAIR — для научных данных: находимость, доступность, совместимость, повторное использование. FAIR не значит «открыто»: набор с закрытыми файлами тоже может им соответствовать, если описание открыто, идентификатор постоянный, а порядок доступа понятен. Городскому CSV они задают словарь, не замену лицензии и выгрузки.

Как проверить свой каталог

Короткий проход по шести пунктам. Подсчёт по выборке из 30 наборов и разбор файлов — в главе о самоаудите.

1. Откройте один файл без платной программы. Если это CSV, прочитайте в карточке кодировку и разделитель, не угадывая.
2. Найдите словарь полей. Если колонка понятна только автору выгрузки, набор для внешнего пользователя не готов.
3. Сверьте дату в карточке с датой внутри файла и с заявленной частотой. Просрочка должна быть видна владельцу раньше, чем пользователю.

4. На карточке есть лицензия ссылкой, издатель, адрес человека, период и территория.
5. Скачайте выгрузку всех описаний без входа. Убедитесь, что в ней тот профиль, который ждёт верхний портал.
6. Откройте старый адрес набора, если платформу уже меняли, и отправьте пробное сообщение об ошибке. Оно должно дойти до владельца данных.

Так выглядит удачный ответ на пятом шаге. Число наборов в постраничной выгрузке data.gouv.fr:

```
curl -s "https://www.data.gouv.fr/api/1/site/catalog.rdf" | grep -o "hydra:totalItems[^<]*"
```

Частота и срок обновления одного набора на HDX:

```
curl -s "https://data.humdata.org/api/3/action/package_show?id=emdat-country-profiles" \
  | jq '{freq: .result.data_update_frequency, due: .result.due_date, period:
.result.dataset_date}'
```

14 октября 2026 года первая команда показывала 74 616 наборов, вторая — частоту `7` (раз в неделю) и срок `2026-10-21`.

Источники

- Data on the Web Best Practices, рекомендация W3C, 31 января 2017 года
- Принципы FAIR; M. D. Wilkinson и др., The FAIR Guiding Principles, Scientific Data, 2016
- Международная хартия открытых данных; число принявших её правительств — с главной страницы, 14 октября 2026 года
- Table Schema
- HDX: поля карточки, как считают свежесть с апреля 2025 года, какую частоту ставить; набор EM-DAT Country Profiles и поиск по частоте `7` — 14 октября 2026 года, 866 наборов
- data.gouv.fr: объявление о балле качества метаданных, 10 августа 2022 года; руководство по баллу; выгрузка DCAT, 74 616 наборов; еженедельные CSV каталога — 14 октября 2026 года
- Объявление о новом каталоге data.gov, 15 апреля 2026 года
- Чек-лист требований, карта возможностей

Антипаттерны

Антипаттерн — решение, которое выглядит как публикация и не даёт данными пользоваться. Ниже те, что видны по живым каталогам. Доли — по выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года), если не сказано иное.

Нет выгрузки описаний

У 34% каталогов не отмечен программный интерфейс. Сборщик федерации и программа, которой нужен один файл всех описаний, такой каталог не заберут. Поисковик страницу откроет, если на ней есть разметка schema.org и карта сайта.

У списков наборов интерфейс отмечен примерно у 14% записей реестра. Находимость такого списка держат лицензия, постоянный адрес и разметка. Список без своего API при десятках наборов — обычный способ, см. когда портал не нужен.

У маркетплейсов интерфейс отмечен у 21% записей, у ML-каталогов — у 23%, у каталогов показателей — у 37%. У показателей это особенно больно: конструктор таблицы в браузере нечем заменить.

Витрина к дате

Наборы выложили один раз и не обновляют. Дата в карточке стоит днём запуска портала. Раздел «Открытые данные» есть, а данных в нём нет: на сайте Росморречфлота в 2026 году страница словами сообщает, что данные публикуются, а ссылка ведёт в никуда, потому что домен портала удалили ещё в 2022 году. Портал Перми в 2023 году заменили страницей с двумя наборами — режимом работы и телефонным справочником, хотя раньше их было больше 480.

Хуже пустой даты только ложная. data.gov.uz показывал в разделе мониторинга обновлённым в декабре 2022 года набор, в котором одна цифра за 2020 год. Отметку об обновлении ставили без изменения данных. Как проверить свой каталог на эти признаки — в главе о самоаудите.

Дробление ради счётчика

Число наборов — самая заметная цифра портала, и её легко накрутить. Один показатель режут на файлы по годам, месяцам и районам, чтобы в отчёте было «несколько тысяч». Пользователю нужен один набор с разрезом или API, а не сто одинаковых имён, которые придётся скачать по одному и склеить.

Как это выглядит в цифрах:

- На портале открытых данных Узбекистана в феврале 2023 года было 6623 набора, из них 42,6% состояли из одной строки, а все данные вместе весили 40 МБ. Летом 2025 года наборов стало 10 412, однострочных — 5207, ровно половина. Больше тысячи записей было только у 114 наборов.

- В разделе открытых данных Росстата осенью 2021 года лежало 1547 наборов, из них 742 — файлы одинаковой структуры, которые можно было выложить одним набором. Без дробления наборов было бы около 190.
- Польский dane.gov.pl в 2020 году заявлял почти 18,8 тысячи наборов, из них больше 7,5 тысячи — помесячные файлы одного патентного ведомства.
- На австрийском data.gv.at в конце 2025 года из 63 тысяч наборов 52 тысячи — бюджеты муниципалитетов, разрезанные по годам.

Родственный приём — набить счётчик чужим содержимым. Вьетнамский data.gov.vn в 2022 году заявлял 10 616 наборов, из них 10 045 — книги и статьи из репозитория университета. Наборов данных было меньше шестисот.

Делить набор можно, если часть нужна сама по себе. Бюджет каждого муниципалитета — отдельный объект: его ищут по названию города. Год того же бюджета — нет, это колонка. Центральные банки вроде ЕЦБ и Банка международных расчётов публикуют каждый временной ряд отдельной записью, и это оправдано: ряд — единица поиска, его так и называют, а весь массив при этом отдаётся одной выгрузкой. Проверочный вопрос простой: станет ли пользователю удобнее, если склеить эти наборы в один. Если да — это дробление. Как считать каталог по существу, а не по числу наборов, — в главе о метриках.

Дашборд вместо данных

Портал показывает графики и виджеты в BI-системе — Power BI, Tableau, Qlik, Superset, — а скачать можно только то, что видно на экране. Национальный портал открытых данных Пакистана nodp.gov.pk в 2026 году — набор дашбордов на Superset со статистикой. Система МАГАТЭ по ядерным реакторам PRIS собрана на Power BI: данные свежие, но получить их можно только экспортом из визуализации, без API. Такие сайты не годятся и для архива: страницы BI-систем веб-архив не сохраняет. Статистический BI-портал Москвы и Московской агломерации на сайте Росстата так и пропадёт без следа, если его выключат.

Дашборд — хорошее дополнение к данным, но не их замена. Новый портал данных МВФ тоже показывает дашборды Power BI, но рядом отдаёт ряды через SDMX API. Порядок такой: сначала набор и выгрузка, потом визуализация.

Файл, который нельзя прочитать

PDF, скан, таблица с объединёнными ячейками и шапкой в три строки. Формат числится машиночитаемым только по расширению. Словаря полей нет.

Крайний случай — документ под видом данных. На сайте МВД России когда-то публиковали как открытые данные XML-файлы, сохранённые из Word: формально XML, по сути разметка текстового документа. Тот же приём — экспорт в «RDF», который на деле обычный XML без словарей и ссылок. Такой файл прибавляет формат в статистике портала и ничего не даёт программе.

Тише, но не менее вредно — файлы без описания технических деталей. В том же разделе открытых данных Росстата (2021) половина CSV была с запятой, половина с точкой с запятой, около 30% в кодировке windows-1251, остальные в UTF-8, и нигде не сказано, где какая. У

части файлов с расширением `.csv` внутри оказывался ZIP. Каждый пользователь угадывает заново.

Права не найти

Лицензию на уровне каталога в реестре удалось записать примерно у 1% записей. Подробности и оговорка — в главе о лицензиях. На практике это значит: даже если лицензия где-то есть, её не видно в описании набора.

Дверь закрыта для программ

Капча, обязательная регистрация, чтобы прочитать описание, блокировка автоматических запросов, лимиты без документации. Метаданные должны читаться без входа. Ограничивать можно файл, не карточку.

Встречаются и более изобретательные запоры. Казахстанский data.egov.kz в 2023 году отменил обязательную регистрацию, но стал отдавать из набора не больше 100 записей за раз; так было и в 2025 году. Европейский каталог масс-спектрометрии Metaspace выдаёт ссылку на скачивание, которая перестаёт работать через полчаса: постоянного адреса у файла нет. С 2024–2026 годов к старым причинам добавилась новая — роботы, собирающие данные для обучения моделей. Патентное ведомство США в 2026 году закрыло массовые выгрузки и API портала data.uspto.gov регистрацией, хотя раньше было одним из самых открытых. Защищаться от нагрузки можно, но так, чтобы программы с понятными намерениями проходили без ручного одобрения: опубликованные лимиты, ключ, который выдаётся автоматически, полная выгрузка файлом, которую не нужно выкачивать запросами.

Та же дверь закрывается целиком, когда сайт не отвечает на запросы из-за рубежа. Многие российские государственные сайты недоступны за пределами российских сетей. Вместе с иностранными пользователями их перестают видеть Internet Archive, поисковики и агрегаторы данных: каталог выпадает и из архива, и из поиска. Проверить это просто: открыть сайт через сервис или сервер в другой стране. Подробнее о последствиях — в главе об архивации.

Свой формат вместо стандарта

Чтобы собрать наборы с тысяч сайтов, придумывают собственный перечень вместо DCAT или schema.org. Так было в России: методические рекомендации обязали ведомства и муниципалитеты вести список наборов в CSV-файле по своей схеме, а data.gov.ru собирал эти списки. Получились тысячи мелких перечней, которые редко обновляли, и формат, который не понимает ни один зарубежный агрегатор и поисковик. Стандарт сбора выбирают из тех, что уже читают другие, — см. федерация и сбор.

Ссылки ломаются при переезде

Платформу сменили, адреса наборов стали другими, перенаправлений нет. Либо файл лежал не в каталоге, а на соседнем сайте, и соседний сайт переехал. Правило сформулировано давно: в заметке W3C «Cool URIs don't change» (1998) Тим Бернерс-Ли пишет, что адрес

меняют не технологии, а люди, которые не подумали о нём заранее. Обратный пример — data.gov: при переезде с CKAN в 2026 году старый каталог оставили по отдельному адресу на переходный срок. Постоянный адрес и план переноса обсуждаются до миграции, см. идентификаторы и стоимость ухода.

Портал бросили

868 каталогов в реестре не действуют, 98 выведены из эксплуатации. Типичная история: пилот на гранте, грант кончился, обновлять некому. Закрытие тоже бывает правильным, если данные и адреса передали. Молчаливое угасание — нет.

Угасание обычно растягивается на годы. Раздел открытых данных Минздрава России наборы с 2017 года не обновлял, а весной 2024 года исчез совсем. Портал Фонда социального страхования data.fss.ru к 2022 году оставался «в тестовом режиме», без ссылки с сайта фонда и с данными за 2018 год. Признаки заброшенности видны задолго до выключения: последняя новость несколько лет назад, план публикации удалён, ссылки на файлы ведут в никуда. См. жизненный цикл и архивацию.

Один набор на трёх порталах

Город, регион и страна показывают одну и ту же карточку и не пишут, кто источник. Поиск врёт о числе наборов, исправление ошибки нужно делать в трёх местах. В карточке копии должна быть ссылка на оригинал.

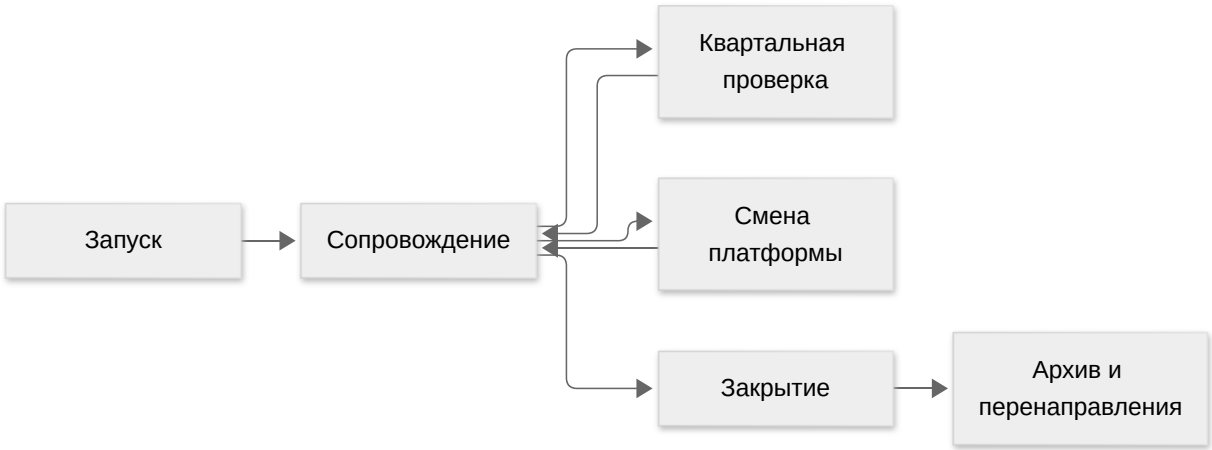
Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года.
- Data on the Web Best Practices, W3C, 2017 — описания, лицензия, постоянные адреса и доступ для программ как обязательные практики
- Cool URIs don't change, W3C, 1998
- Объявление о новом каталоге data.gov, 15 апреля 2026 года
- И. Бегтин, Что не так с порталом открытых данных Узбекистана?, февраль 2023 — размеры наборов и мониторинг обновлений на data.gov.uz
- И. Бегтин, Почему невозможно хвалить Росстат, январь 2022 — разбор раздела открытых данных Росстата
- И. Бегтин, Хорошие и плохие практики публикации данных, ноябрь 2024 — пример МВД; каталоги плохих практик OpenDataFail и Bad Data
- И. Бегтин, Почему данные по России так непросто найти, ноябрь 2024 — недоступность сайтов из-за рубежа и собственный формат сбора
- И. Бегтин, Telegram, о дроблении: dane.gov.pl, 17.12.2020; data.gov.vn, 21.08.2022; data.egov.uz в 2025 году, 10.07.2025; data.gv.at, 23.12.2025; когда нарезка оправдана, 11.01.2024
- И. Бегтин, Telegram, о дашбордах: PRIS, 11.09.2026; nodp.gov.pk, 15.08.2026; BI-портал Росстата, 23.05.2025; data.imf.org, 19.04.2025

- И. Бегтин, Telegram, о закрытом доступе: data.egov.kz, 26.04.2023, и в 2025 году, 29.10.2025; Metaspace, 11.07.2024; data.uspto.gov, 23.06.2026
- И. Бегтин, Telegram, о заброшенных разделах: Пермь и Приморье, 02.04.2023; Минздрав, 13.11.2024; Росморречфлот, 13.05.2026; ФСС, 15.07.2022
- Хорошие практики

Жизненный цикл каталога

Каталог запускают за недели и сопровождают годами. Большинство сбоев — не в день открытия, а когда кончился проект, сменилась платформа или ушёл единственный администратор.



Запуск

Запуск — день, когда посторонний человек открывает настоящий набор, скачивает файл и видит, кто обновит его в следующем году. Публичный адрес — последний шаг этого дня. Срок установки разных способов — в главе о развёртывании.

Пилот — около десяти настоящих наборов, и среди них есть неудобный. Неудобный набор — тот, у которого в источнике нет лицензии, файл большой или обновление обещано каждый день. На пилоте видно, кто допишет лицензию, кто заметит сломанную ссылку и читает ли верхний портал вашу выгрузку. Демонстрационные наборы поставщика на эти вопросы не отвечают. Что проверить на стенде — в главе о выборе платформы.

За данные и за платформу отвечают разные люди. Их называют до публичного адреса, по именам.

Так это устроено у Нью-Йорка. Портал nyc.gov/opendata ведёт команда открытых данных в Управлении технологий и инноваций: стратегия, стандарты, сам сайт. Каждое городское ведомство назначает своего координатора открытых данных. Через него ведомство отдаёт наборы на портал, обновляет их с той частотой, которая записана в карточке, отвечает на вопросы жителей и готовит свою часть годового отчёта. Если координатор уходит, ведомство обязано назначить другого и сообщить об этом команде портала. Это записано в руководстве по техническим стандартам, редакция 12 мая 2023 года. Схема живая: рабочая тетрадь годового отчёта за 2025 год на сайте портала обновлена 23 июня 2025 года.

Перед публичным адресом:

- [ ] Пилот прошёл на настоящих наборах, включая неудобный.
- [ ] Названы два человека: кто обновляет строки и кто обновляет платформу.

- [ ] Схема адресов записана так, чтобы пережить следующую платформу. Как её удержать — в главе об идентификаторах.
- [ ] Полную выгрузку описаний и файлов один раз скачали и открыли.
- [ ] На следующий год названы ставка и бюджет. Подрядчик, который уходит в день акта, эти места не занимает. Что записать в договор — в главе о способах создания.

Открытие сайта бывает без запуска. 15 июля 2025 года Минэкономразвития снова открыло data.gov.ru. Прежнюю версию закрыли в марте 2023 года «для переноса на новую платформу». Сайт заработал. Через API на нём было 5826 наборов, около 100 МБ в сжатом виде, и 84% из них впервые опубликованы в 2019 году или раньше. Снимок прежней версии от 2 февраля 2022 года — 26 490 наборов, архив 14 ГБ — на новый портал не вернулся. Он лежит в Национальном цифровом архиве, см. ниже. Разбор по числу строк и по тому, каких ведомств на портале нет, — в главе о самоаудите. Запуском такое открытие становится, когда на новом адресе появляются свежие наборы с названным владельцем.

Сопровождение

Дальше каталог живёт проверками. Число наборов в отчёте их не заменяет.

Календарь ссылок и полного прохода уже задан в главе о самоаудите: раз в месяц открывают десяток известных ссылок, раз в полгода проходят выборку целиком. К этому имеет смысл добавить две более редкие проверки. Их делают те же двое, что названы при запуске: строки и карточки смотрит тот, кто отвечает за данные, версию, копии и счёт — тот, кто отвечает за платформу.

Как часто	Что сделать
Раз в квартал	Лицензия на наборе заполнена. Дата в файле совпадает с обещанной частотой. Версия платформы ещё поддерживается. Оба человека из запуска на месте.
Раз в год	Копию пробуют развернуть. Полную выгрузку снова скачивают. Дата окончания лицензии или контракта на коммерческую платформу стоит в календаре.

Ссылки. Если месячную проверку пропускают годами, получается британский data.gov.uk. С 2017 года портал недофинансировали. К марту 2026 года не меньше четверти ссылок вели на страницу с ошибкой, а многие наборы годами не обновляли организации, которых уже нет. Платформу при этом оставили: прежний каталог работает, а рядом собрали витрину отобранных коллекций. Как устроена эта витрина — в разделе о смене платформы.

Частота. Дата в карточке — дата строк. На Humanitarian Data Exchange письмо о просрочке уходит ответственному за набор, а не в общий ящик портала. Как считают срок — в главе о хороших практиках. Своему каталогу достаточно того же адресата.

Версия. Платформу и расширения обновляют, пока версия ещё в окне поддержки. У CKAN полностью поддерживается последняя минорная версия, предыдущая получает исправления ошибок и безопасности. 14 октября 2026 года это 2.12 и 2.11. Выпуск старше предыдущего уже вне окна, и обновление обычно превращается в перевнедрение: по дороге меняются расширения, схема и способ хранения файлов. Окна других платформ — в главе о развёртывании.

Люди. Сайт может открываться, когда отвечать уже некому. В Нью-Йорке замена координатора — обязанность ведомства. Другой конец той же истории виден по реестру. По выгрузке реестра dataportals-registry (выпуск 1.27.0 от 15 октября 2026 года) не действуют 868 каталогов, выведены из эксплуатации 98. Доля неработающих сильно различается: около 0,9% у каталогов показателей и около 18% у каталогов микроданных. За второй цифрой стоят в основном брошенные установки: 47 из 59 неработающих каталогов микроданных — старые каталоги NADA статистических служб, часто на IP-адресах и без HTTPS. Их ставили в рамках программ поддержки, а сопровождение в бюджет не заложили. Подробности — в главе о каталогах микроданных. Каталог без владельца и бюджета на сопровождение умирает вместе с проектом. Признаки, которые снаружи видны за годы до выключения, собраны в антипаттерне «Портал бросили».

Годовая проверка копии — та же мысль, что в главе о развёртывании: копия, которую никто не пробовал развернуть, копией не считается. Дата лицензии ArcGIS, на которой остановился геопортал Архангельской области, показывает, зачем эта дата стоит в календаре заранее. Сам случай — среди провалов переезда.

Смена платформы

Смена — главная причина мёртвых ссылок. До переноса делают выгрузку всех описаний и файлов и таблицу старых и новых адресов. После переноса старые адреса перенаправляют. Если перенаправить всё сразу нельзя, старую систему оставляют по отдельному адресу на переходный срок. Так в апреле 2026 года сделал data.gov: каталог переехал с CKAN на собственную разработку, страницы наборов остались по прежним адресам, а старый сайт оставили на catalog-old.data.gov. В описании проекта, обновлённом 23 июля 2026 года, этот адрес ещё числился работающим всю осень. 14 октября 2026 года сайт уже не открывается: имя catalog-old.data.gov ещё есть, но указывает на адрес, которого нет. Прежний программный интерфейс CKAN по старому пути отвечает ошибкой. Новый интерфейс живёт по другому адресу и просит ключ, так что программы, которые ходили в CKAN, пришлось переписывать. Как оценить цену ухода заранее, написано в главе о способах создания.

Переезжают в обе стороны, и ни одно направление само по себе не ошибка:

- **На CKAN.** Австралийский data.gov.au с 2019 года жил на смешанной схеме: поиск и витрину давала Magda, под ней стоял устаревший CKAN. В июне 2025 года портал свели к одному CKAN с Drupal для страниц, сохранили больше 100 тысяч наборов и сбор с тысячи источников, а программный интерфейс CKAN остался прежним. Греческий data.gov.gr в 2025–2026 годах перешёл на CKAN со своей разработки и за счёт сбора чужих каталогов вырос с сотни наборов до десятков тысяч. 14 октября 2026 года это CKAN 2.11.3 и 22 762 набора.
- **С CKAN.** Кроме data.gov, на собственный движок ушёл монгольский opendata.gov.mn, и вместе со старой платформой пропал программный интерфейс. 14 октября 2026 года путь прежнего API CKAN открывает страницу сайта, а не ответ интерфейса.
- **Без смены платформы.** Британский data.gov.uk в 2026 году стал National Data Library: новая витрина с отобранными коллекциями, а прежний каталог оставлен отдельным разделом, чтобы было с чем сравнивать.

Типичные провалы тоже повторяются. Словацкий портал при переезде на data.slovensko.sk в 2024 году сменил адреса наборов, и ссылки на них из европейского портала перестали работать. Малайзия, переводя data.gov.my на новый движок, оставила старый каталог по адресу archive.data.gov.my. 14 октября 2026 года соединение с этим адресом не устанавливается, так что временный архив оказался временным. Геопортал Архангельской области потерял слои, когда кончилась лицензия ArcGIS: каталог продолжал показывать карточки слоёв, которых больше не было. Российский data.gov.ru в 2023 году закрыли «для переноса на новую платформу» раньше, чем она появилась, и два года портала не было вовсе.

Отсюда три правила. Старую версию держат, пока новая не приняла все адреса. Программный интерфейс — такая же часть адреса, как страница набора: если он меняется, сборщики теряют каталог. Лицензия на коммерческую платформу — дата, к которой план переезда должен быть готов.

Закрытие

Закрыть каталог нормально, если им никто не пользуется или платформу больше не на чем держать. Плохо закрывать его выключением сервера в тот день, когда об этом узнали пользователи.

До выключения каталог несколько месяцев работает только на чтение. Сайт, файлы и программный интерфейс открываются, новые правки остановлены, на первой странице написано, почему так и до какой даты. Пауза без даты — это уже заброшенный сайт. data.gov.ru эту стадию пропустил: портал выключили «на время переноса», и два года по старому адресу не было ни чтения, ни объяснения, куда делись наборы.

Перед выключением:

- [ ] О закрытии объявлено заранее, за месяцы. До объявленной даты каталог работает только на чтение.
- [ ] Данные переданы в другой каталог или архив, либо на старом адресе прямо сказано, что они сняты с публикации.
- [ ] Старые адреса перенаправляют. Если перенаправить всё нельзя, хотя бы год они отвечают, куда переехал набор.
- [ ] Снимок описаний сохранён: выгрузка через программный интерфейс или файл DCAT, а не только страницы сайта.
- [ ] Файлы, которые лежат только здесь, сохранены вместе с адресами, откуда они взяты.
- [ ] Сайт есть в Интернет-архиве или национальном веб-архиве, и архив до него доходит, в том числе из-за рубежа.
- [ ] Верхний портал и реестры, где каталог был зарегистрирован, уведомлены.

Тот же список — в чек-листах.

Как снять снимок

«Снимок описаний сохранён» — это не скриншот главной страницы. Архивация каталога проходит четыре шага.

1. **Найти всё, что относится к каталогу.** Основной сайт, поддомены, старые версии вроде `old.` и `archive.`, отдельные сервисы карт и API. Старые версии сайта исчезают первыми: их держат «на всякий случай», а потом выключают незаметно.
2. **Расставить приоритеты.** Сначала то, что пропадёт раньше и чего нет больше нигде: файлы, которые лежат только здесь, сервисы на выключаемой платформе.
3. **Собрать.** Страницы сайта — веб-архивом в формате WARC, который читают Internet Archive и национальные библиотеки. Описания — выгрузкой через API или файлом DCAT. Файлы — скачать как есть, вместе со ссылками, откуда они взяты. Если данные отдаются только через API, нужна отдельная выгрузка, например инструментом APIBackuper: веб-архиватор API не обходит.
4. **Опубликовать или сдать на хранение.** Файлы WARC и выгрузки кладут в открытое хранилище, а их описание — в каталог, чтобы архив можно было найти. Минимум — сдать копию в холодное хранилище и записать, где она.

Посмотреть WARC без сервера можно в ReplayWeb.page. Для сайтов с динамической подгрузкой нужен краулер через браузер, например Browsertrix Crawler.

Если закрывают не вы

Каталог могут закрыть те, кто им владеет, не спросив пользователей. Тогда данные переживают его только в чужих архивах. Российский data.gov.ru закрыли в марте 2023 года «для переноса на новую платформу». Полный снимок прежней версии — 26 490 наборов, 14 ГБ в сжатом виде — сохранился в Национальном цифровом архиве: его сделали 2 февраля 2022 года. Версия, открытая 15 июля 2025 года, этих данных не вернула. Как исчезают каталоги, почему веб-архив сохраняет их плохо и как снять копию чужого каталога — в главе «Архивация и спасение каталогов».

Каталоги, которые просто перестали сопровождать, посчитаны в разделе о сопровождении. Отдельный случай — когда снимают с поддержки сам продукт. Nesstar в 2015 году с версией 4.0 объявили устаревшим, к 2022 году окончательно сняли. Канадский архив социальных обследований ODESI выбрал для переезда Borealis на Dataverse, где есть описания DDI и DOI, и выключил Nesstar только после переноса: в октябре 2026 года старый адрес ведёт на новый сайт. Закрытие чужого продукта встречают планом миграции, пока старая система ещё открывается.

Источники

- Выгрузка реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года.
- Координаторы открытых данных Нью-Йорка — обязанности координатора; рабочая тетрадь годового отчёта за 2025 год обновлена 23 июня 2025 года
- Technical Standards Manual, редакция 12 мая 2023 года — кто ведёт портал, кто ведёт наборы ведомства, замена координатора при уходе
- Объявление о новом каталоге data.gov, 15 апреля 2026 года; описание каталога, обновлено 23 июля 2026 года — старый адрес числился работающим всю осень; новый программный интерфейс — другой адрес и ключ. Недоступность catalog-old.data.gov и ошибка старого Action API — проверка 14 октября 2026 года

- Link Digital, перевод data.gov.au на CKAN, 2025; GDS, What's changing on data.gov.uk and why, 25 марта 2026 года — недофинансирование с 2017 года, не меньше четверти ссылок ведут на ошибку
- И. Бегтин, Telegram: data.gov.gr на CKAN, 18.05.2026; opendata.gov.mn, 25.08.2026; data.slovensko.sk, 20.06.2024; data.gov.my, 19.01.2024; геопортал Архангельской области, 21.08.2024
- Nesstar: история продукта, перенос ODESI на Borealis
- Формат WARC в описании Библиотеки Конгресса, ReplayWeb.page, Browsertrix Crawler, APIBackuper — выгрузка данных из API для архива
- И. Бегтин, Как правильно архивировать цифровые ресурсы при архивационных кампаниях, ноябрь 2021 — четыре стадии архивации; Веб-архивация и её ограничения, февраль 2025
- И. Бегтин, Обновлённый портал data.gov.ru, 16 июля 2025 года — 5826 наборов, около 100 МБ, 84% с первой публикацией в 2019 году или раньше; архивная копия от 2 февраля 2022 года
- Хорошие практики, самоаудит, каталоги микроданных
- data.gov.gr: `status_show` и `package_search` 14 октября 2026 года — CKAN 2.11.3, 22 762 набора. Путь API opendata.gov.mn и адрес archive.data.gov.my проверены в тот же день.
- Политика поддержки версий CKAN — 14 октября 2026 года полностью поддерживается 2.12, исправления для 2.11
- Скорость развёртывания и сопровождение
- Постоянные идентификаторы

Архивация и спасение каталогов

Каталог исчезает чаще, чем кажется, и обычно без объявления. В главе о жизненном цикле закрытие описано глазами владельца: что сделать до выключения сервера. Эта глава — о другой стороне. Владелец не позаботился об архиве, решение о закрытии принято не им, или вы пользуетесь чужим каталогом и не уверены, что он доживёт до следующего года.

Главный вывод: веб-архив не спасает каталог сам по себе. Страницы он сохраняет, а описания из программного интерфейса, большие файлы и сайты, закрытые для иностранных адресов, — плохо или никак. Копию нужных наборов делают заранее и целенаправленно.

Как исчезают данные

В 2017 году Sunlight Foundation описала восемь способов, которыми из государственных сайтов США пропадает информация. Для каталогов важны три: удаление наборов данных, удаление или переделка поисковых порталов и перенос страниц без перенаправлений. На практике это выглядит так.

Тихое удаление. Росимущество в 2023 году убрало раздел открытых данных: 24 набора, около 1 ГБ XML. Ссылки на раздел исчезли со всех страниц сайта, объявления не было. Когда это случилось, удалось установить только по Интернет-архиву: последний снимок страницы сделан в феврале 2022 года.

Выключение решением за несколько дней. Портал открытых данных Московской области data.mosreg.ru вывели из эксплуатации в марте 2019 года. Распоряжение подписали 25 марта, систему выключили 29-го, а само распоряжение опубликовали через два месяца. Архив портала никто не сделал.

«Временно, на время переезда». Российский data.gov.ru закрыли в марте 2023 года для переноса на новую платформу. Новая версия открылась только летом 2025 года и прежние данные не вернула. Подробнее — в главе о жизненном цикле.

Ликвидация владельца. При реорганизации архив переходит к правопреемнику, при ликвидации — никуда. Сайт Уставного суда Санкт-Петербурга закрыли одновременно с ликвидацией суда в 2021 году. При упразднении Ростуризма в 2022 году под угрозой оказались его поддомены, включая портал открытых данных.

Прекращение публикации вместе с архивом. Ведомство решает больше не выпускать ряд и заодно удаляет всё, что выпустило раньше. Так в 2024 году Банк России поступил со статистикой внебиржевого валютного рынка, а Росавиация убрала архив статистики перевозок с 2010 года.

Смена политики. В США в 2025 году с федеральных сайтов массово удаляли данные о климате, здоровье, международной помощи; сайт USAID отключили вместе с порталом данных агентства. Хронологию удалений ведёт Архив национальной безопасности при Университете Джорджа Вашингтона.

Закрытие чужой площадки. Коммерческие и общие хабы тоже закрываются: в 2024 году Microsoft закрыла облачную среду Planetary Computer Hub, где исследователи работали с геоданными, и пользователям пришлось срочно забирать блокноты и файлы. Риски хабов разобраны в главе об облачных хранилищах.

Во всех случаях первым признаком становится не объявление, а пропавшая ссылка.

Почему веб-архив не спасает каталог

Интернет-архив и национальные веб-архивы обходят сайты краулером: открывают страницу, сохраняют её и идут по ссылкам дальше. Каталог данных устроен так, что этот способ многое теряет.

- **Интерфейс поверх API.** Список наборов подгружается скриптом из программного интерфейса, у набора нет отдельной страницы со ссылкой на файл. Краулер видит пустую оболочку. Так был устроен единственный известный портал открытых данных Афганистана на GeoNode: в декабре 2022 года он исчез, а в веб-архиве от него осталась главная страница. Архив бюллетеней Счётной палаты России с 1996 года тоже отдаёт документы через поисковый API, и до самих PDF краулеры не доходят.
- **Большие файлы и архивы.** Интернет-архив редко сохраняет ZIP и многогигабайтные файлы, а страницы, на которые мало ссылок, обходит реже.
- **Закрытость для иностранных адресов.** С 2022 года многие российские государственные сайты не отвечают на запросы из-за рубежа. В 2023 году сайты Минюста и МВД в Интернет-архиве не обновлялись вовсе, сайт Минсельхоза не сохранялся около трёх лет. Если архив не может открыть сайт, копии не будет, даже если сайт работает.
- **Запрет в robots.txt.** Владелец может сам закрыть сайт от архива. «Карта российской науки» так и исчезла в 2017 году: в Интернет-архиве сохранилась только главная страница.
- **Один архив на всех.** Почти весь мир полагается на Интернет-архив, у которого нет устойчивого финансирования. В 2025 году у него был многомесячный сбой, и снимки многих сайтов не делались. Часть издателей в 2026 году стала блокировать его краулер, опасаясь, что через архив тексты заберут компании, обучающие модели.

Поэтому копия каталога — это не только снимок страниц в формате WARC. Это ещё выгрузка описаний через API или файл DCAT и скачанные файлы. Как это делает владелец перед закрытием, описано в разделе «Как снять снимок» главы о жизненном цикле. Тот же порядок годится и для чужого каталога.

Как бывает правильно

Национальный архив хранит сайты государства. В Великобритании веб-архив Национального архива регулярно сохраняет сайты всех государственных органов. Когда статистическая служба ONS переделала сайт и сменила систему хранения рядов, адреса изменились, но данные не пропали: страница инструмента временных рядов ведёт на копию старого сайта в архиве. Пользователь, у которого в статье стоит старая ссылка, находит то, на что ссылался.

Закрытие с предупреждением. Microsoft в марте 2017 года объявила о закрытии хостинга кода CodePlex: с октября сайт работал только на чтение, в декабре закрылся, к закрытию выпустили руководство по переносу. Девять месяцев хватило, чтобы владельцы проектов и волонтёры сохранили всё нужное. В том же году российский сервис рекомендаций Имхонет закрылся за день, без возможности забрать свои материалы.

Копия с доказательством происхождения. Библиотечная лаборатория Гарвардской школы права в 2024–2025 годах сохранила всё, на что ссылался data.gov: около 311 тысяч наборов, 17,9 ТБ. Каждый набор упакован по стандарту BagIt вместе с метаданными из API и цифровой подписью с меткой времени, так что любой может проверить, что файл не изменён после сохранения. Архив лежит в открытом облачном хранилище Source Cooperative, к нему есть поиск, а код для таких архивов открыт.

Даже большие сервисы отступают. Google в 2024 году объявил, что с 25 августа 2025 года перестанут работать все короткие ссылки goo.gl. Волонтёры ArchiveTeam начали обходить их заранее. За несколько недель до срока Google изменил решение и сохранил ссылки, по которым ещё переходили, а отключил только неактивные. Короткая ссылка не постоянный идентификатор: её судьбу решает владелец сервиса, а не тот, кто её поставил в документ. Подробнее — в главе об идентификаторах.

Кто спасает, когда владелец не стал

Государственные и университетские программы. Проект End of Term Web Archive с 2008 года сохраняет сайты федерального правительства США при каждой смене президента. Архив 2024–2025 годов собрали шесть организаций: Интернет-архив, ArchiveTeam, Common Crawl, гарвардская лаборатория, библиотеки Университета Северного Техаса и Webrecorder. Это 2,29 ПБ сжатых файлов WARC. Они лежат в открытом хранилище AWS под лицензией CC0, сложные интерактивные сайты отдельно доступны для просмотра на GovArchive.us.

Сообщество. ArchiveTeam — добровольцы, которые сохраняют закрывающиеся сервисы распределённым краулером. Data Rescue Project в 2025 году объединил исследовательские организации, которые сохраняют именно данные, а не страницы, и ведёт открытый каталог спасённого. Ещё раньше, в 2017 году, по тем же причинам появились DataRefuge и Climate Mirror.

В России этим занимается Национальный цифровой архив (ruarxive.org) — общественный проект, а не государственная служба. В 2022 году он провёл кампанию архивации порталов открытых данных: к маю собрал 37 сайтов, 43 ГБ в сжатом виде; самые большие слепки — порталы Минкультуры и data.gov.ru. В 2025 году, когда объявили о выводе из эксплуатации системы статистики ЕМИСС, архив сохранил сайты Росстата, его территориальных органов и саму ЕМИСС. Обязательной сдачи сайтов и данных государственных органов в архив в России нет.

У всех этих архивов одна слабость: ими можно воспользоваться, только если о них знают. Копия, о которой нет записи ни в одном каталоге, почти так же недоступна, как удалённый оригинал. Поэтому архивы публикуют вместе с описаниями — в том же формате, что и каталог данных.

Как сохранить каталог, которым вы пользуетесь

1. **Проверьте, есть ли внешняя копия.** Откройте адрес набора в Интернет-архиве и посмотрите, когда был последний снимок и сохранились ли файлы, а не только страница. Проверьте сайт и из другой страны: если он закрыт для иностранных адресов, внешней копии почти наверняка нет.
2. **Снимите описания целиком.** У CKAN это постраничный обход `/api/3/action/package_search`, у многих порталов — общий файл DCAT или `/data.json`, у научных репозиториев — OAI-PMH. Описания весят мало, а без них архив файлов превращается в папку с непонятными именами.
3. **Скачайте файлы, которые вам нужны.** Сохраните рядом адрес, с которого файл взят, дату и контрольную сумму. Если данные отдаются только через API, обычный архиватор их не возьмёт: нужна выгрузка постранично, например инструментом APIBackuper.
4. **Упакуйте с описанием.** Для набора файлов подходят BagIt или Frictionless Data Package: в пакете лежат и данные, и манифест. Для страниц сайта — WARC или его упакованный вариант WACZ, который открывается в ReplayWeb.page без сервера.
5. **Решите, можно ли публиковать копию.** Открытая лицензия на наборы разрешает выложить их копию в открытое хранилище и описать её в каталоге. Но она касается только того, на что её дал владелец. Чужие файлы, на которые портал лишь ссылается, персональные данные и условия использования самого сайта она не покрывает. Копия всего каталога целиком может задеть ещё и особое право на саму базу данных: в ЕС и России оно есть у того, кто вложился в её создание. Такие части проверяют отдельно. Если лицензии нет или она запрещает распространение, копию держат у себя и сообщают о ней архиву или владельцу.

Что из этого следует для владельца

Если вы ведёте каталог, проще всего не дать ему стать предметом спасательной операции:

- не закрывайте сайт от архивных краулеров и от иностранных адресов, если на нём открытые данные;
- держите полную выгрузку описаний и файлов по постоянному адресу: её заберут и агрегатор, и архив;
- прекращая выпуск ряда, не удаляйте то, что уже опубликовано;
- о закрытии объявляйте заранее, на месяцы, а не на дни, и оставляйте режим «только чтение»;
- если в стране есть национальный архив сайтов, договоритесь о сдаче копии, как это делают британские ведомства.

Пункты для чек-листа — в разделе «Перед выключением каталога».

Источники

- Sunlight Foundation, Classifying changes to public access to information on US government websites, октябрь 2017 года (копия в Интернет-архиве)
- Хронология удаления данных в США, National Security Archive, 30 марта 2026 года

- End of Term 2024 Dataset — состав и объём; End of Term Web Archive в реестре открытых данных AWS; GovArchive.us
- Harvard Library Innovation Lab, Announcing the Data.gov Archive, 6 февраля 2025 года; поиск по архиву; архив на Source Cooperative
- Data Rescue Project, ArchiveTeam
- Google, обновлённые планы для ссылок goo.gl, 1 августа 2025 года
- Веб-архив государственных сайтов Великобритании; инструмент временных рядов ONS со ссылкой на архивную копию
- Национальный цифровой архив; код кампании 2022 года — rudatarchive
- BagIt, RFC 8493, WACZ, ReplayWeb.page, APIBackuper
- И. Бегтин, Telegram: data.mosreg.ru, 03.06.2019; Sunlight Foundation о способах исчезновения информации, 09.10.2017; CodePlex и Имхонет, 29.04.2017; что не поддаётся веб-архивации, 25.12.2020; кампания архивации порталов, 20.05.2022; архив Счётной палаты, 30.11.2022; Росимущество, 17.02.2023; портал Афганистана, 05.07.2023; госсайты и Интернет-архив, 19.12.2023; ONS и Национальный архив, 14.02.2024; архивация Росстата и ЕМИСС, 19.05.2025; сбой Интернет-архива, 24.10.2025
- Жизненный цикл каталога, постоянные идентификаторы

ЧАСТЬ X.

Качество каталогов

Уровни качества, метрики, методики и рейтинги, самоаудит каталога данных.

1. Уровни качества
2. Метрики качества
3. Методики и рейтинги
4. Самоаудит каталога

Уровни качества

«Качество» в задаче называют уровнем. Их четыре, и высокий балл на одном оставляет остальные как были. Прежде чем чинить или отчитываться, спросите, какой уровень трогает оценка, и поручите работу тому, кто этот уровень держит.

Число наборов, индекс страны и балл карточки отвечают на разные вопросы. Как считать доли — в главе о метриках. Что меряет каждая известная методика — в главе о методиках. Как пройти свой каталог за один раз — в главе о самоаудите. Здесь — как разложить оценку по уровням.

Четыре уровня

Уровень	Вопрос	Чем обычно меряют	Кто чинит	Что остаётся непроверенным
Данные	можно ли пользоваться файлом или слоем	формат, полнота, совпадение с источником, повторяемость расчёта	владелец данных	заполненность карточки, доступность сайта, индекс страны
Описания	можно ли найти набор и понять условия	заполненность полей, соответствие профилю, живая ссылка	публикатор и редактор каталога	число строк в файле, политика страны
Каталог как сервис	открывается ли он и отдают ли его машины	доступность, API, выгрузка, лицензии в выгрузке, постоянные адреса	администратор платформы	содержимое файлов, индекс страны
Система каталогов	есть ли цепочка и доходят ли по ней лицензия и источник	индексы открытости, доля наборов источника на верхнем портале, дубли	координатор открытых данных	один городской файл

Пустой файл с заполненной карточкой — годные описания и пустые данные. Живой сайт без выгрузки — годная витрина и слабый сервис. Страна с сильным национальным порталом и мёртвыми городскими — сильный верх и слабая система. Бывает и наоборот: файл и карточка у города в порядке, а на верхнем портале лицензия пропадает. Так ушёл набор города Мальмё в 2024 году; разбор цепочки — в главе о федерации.

Один объект, разные баллы

Портал выглядит живым, файлы пустые. На data.gov.uz в феврале 2023 года было 6623 набора — больше, чем на большинстве порталов региона. Раздел мониторинга показывал почти все наборы обновлёнными. По уровню сервиса это выглядело живым. По уровню данных — нет: 96% наборов меньше 200 строк, 42,6% — одна строка, весь портал в CSV занимал 40 МБ. Карточки при этом могли быть заполнены. Балл описаний видит заполненные поля и оставляет пустые строки незамеченными: их видно после открытия файлов. Как считать распределение по строкам и второй пример, data.gov.ru в 2025 году, — в главе о самоаудите.

Значок есть, расчёт не повторяется. Воспроизводимость — возможность заново получить заявленный результат по выложенным данным и записанным шагам. Значок «открытые данные» у журнала говорит, что файл выложен. В 2021 году Хардвик и соавторы проверили 25 статей Psychological Science с таким значком за 2014–2015 годы. Без помощи авторов

проверенные числа сошлись у 9 статей из 25. После пояснений авторов — у 15 из 25. Главная причина расхождения в самой работе — шаги анализа записаны неясно, хотя значок уже подтверждал, что данные выложены. Совпадает ли расчёт с текстом статьи, проверяют отдельно.

Балл карточки смотрит на карточку. data.gouv.fr показывает издателю балл качества метаданных: описание, файлы, частота, лицензия, открытый формат, территория, период. Пустой файл с заполненными полями получает высокий балл. Что входит в балл и как он выглядел 14 октября 2026 года — в главе о хороших практиках. Это уровень описаний.

Что считать годным

Годный ответ зависит от типа каталога. Шесть практик, без которых набор нельзя найти или использовать, — в главе о хороших практиках. Ниже — чем «данные» и «описания» отличаются у соседей и где уровень ломается первым.

Тип	Данные годны, когда	Описания годны, когда	Где обычно ломается первым
Портал открытых данных	файл открывается программой, в нём есть строки и словарь колонок	лицензия, издатель, период и территория стоят на наборе	счётчик растёт за счёт пустых и нарезанных таблиц
Геопортал	слой скачивается файлом или через сервис WFS	у слоя есть лицензия, дата и описание	муниципальная карта: слой виден, кнопки скачивания нет
Научный репозиторий	по файлам и записанным шагам расчёт повторяется	есть DOI, авторы, лицензия и ссылка для статьи	значок и DOI есть, шаги анализа не записаны
Каталог показателей	ряд продолжается и его можно забрать целиком	методика лежит рядом с рядом, коды взяты из общего справочника	API отвечает, а ряд остановился
Каталог микроданных	по словарю переменных файл можно прочитать	описание открыто, уровень доступа назван	файл закрыли вместе с карточкой
ML-каталог	файл скачивается, деление на обучающую и тестовую части неизменно	в карточке есть лицензия и происхождение	карточка заполнена, лицензия свободнее, чем у исходных данных
Каталог метаданных	запись словаря или кода полна и указывает на действующую версию	запись находится по коду и теме	карточка жива, а каталог, на который она указывает, выключен
Страница, Git или хаб	файл открывается без входа, кодировка и колонки написаны	есть README, лицензия, разметка schema.org, дата и контакт	репозиторий есть, страницы с постоянным адресом нет

У микроданных закрытый файл при открытой карточке — нормальный уровень, если порядок доступа написан. Ломается связка, когда прячут и описание.

Разбор поломок — в главах о порталах открытых данных, геопорталах, научных репозиториях, показателях, микроданных, ML-каталогах и каталогах метаданных.

Проверка на входе

Всё, что выше, меряют после публикации. Можно и до: каталог не принимает набор, который не прошёл проверку. Тогда качество — условие входа, а не показатель на дашборде.

Так работают архивы нейроданных. OpenNeuro принимает только наборы по стандарту BIDS — общему порядку папок, имён файлов и описаний. С предупреждениями валидатора загрузить можно, с ошибками нельзя. В DANDI кнопка публикации становится активной только после того, как исправлены все ошибки проверки описания и файлов. Опубликованная версия не меняется.

Проверка на входе работает там, где у области есть общий стандарт. Иначе она становится списком правил одного портала, и издатели обходят его пустыми полями. Для таблиц на портале открытых данных аналог — проверка файла по схеме при загрузке. Схемы для типовых таблиц уже собраны в каталоге schema.data.gouv.fr. Уровень, который такая проверка держит, — данные. Чинить файл после неё по-прежнему владельцу данных, но уже до публикации, а не после жалобы.

Другие новшества того же ряда — в главе о передовых каталогах.

Куда класть чужой балл

Чужую оценку сначала кладут на один уровень. Потом уже решают, чинить ли по ней свой каталог.

Оценка	Какой уровень она трогает
Пять звёзд	данные: формат одного набора
FAIR и F-UJI	описания одной научной записи и порядок доступа
MQA, балл data.gouv.fr	описания каталогов, которые уже отдают карточки наверх, и живы ли ссылки
CoreTrustSeal	каталог как архив: процедуры и сохранность
Open Data Maturity, ODIN, Global Data Barometer	система каталогов и политика страны
Число наборов	счётчик карточек
Балл доверия реестра	грубый вид снаружи: жив ли сайт, указана ли лицензия

MQA видит каталоги, которые уже отдают описания в data.europa.eu, и усредняет те карточки, которые оценка успела посчитать: это может быть доля каталога в поиске. Индекс страны складывается из политики и национальных порталов. Формулы и границы каждой методики — в главе о методиках.

В каком порядке чинить

Снизу вверх по зависимостям. Сначала файлы, которые открываются и читаются. Затем лицензия и обязательные поля. Затем выгрузка и постоянные адреса: без них описания не дойдут до верхнего портала и поисковиков. Индекс страны сдвигается, когда нижние уровни чинят во многих каталогах сразу.

Срок сбора сверху двигает выгрузку вперёд очереди — для тех наборов, которые вы правда публикуете. Заполнять карточки однострочных таблиц к этому сроку значит чинить описания там, где сломаны данные.

Задачу ставят на том уровне, где сломано. Редактору карточек поручают пустые поля. Владельцу данных — файл и его свежесть. Администратору платформы — выгрузку, API и адреса. Координатору страны — правила, по которым города отдают описания, и ответы для индекса вроде Open Data Maturity: этот балл собирают по политике и национальным порталам. Один CSV остаётся у владельца данных.

Пример: ведомству сказали поднять качество

Контекст. Региональный портал открытых данных. В мониторинге почти все наборы помечены обновлёнными. Национальный индекс страны средний, и от региона ждут вклад. Пользователь прислал бюджет: таблица из одной строки в PDF. Редакция — два человека.

Варианты. Заполнить пустые поля у всех карточек. Переписать API и выгрузку. Готовить ответы для странового индекса. Открыть выборку файлов, починить формат и состав у наборов, которыми пользуются, затем лицензию и обязательные поля, затем выгрузку.

Критерии. Человек открывает файл и пользуется им без ручной правки. Программа забирает описания без входа. Работа попадает к тому, кто держит уровень. Два редактора успевают за квартал.

Решение. Месяц уходит на данные тех наборов, которые есть в письмах и в статистике скачиваний: открытый формат, строки, словарь колонок. Следующим шагом — лицензия и обязательные поля на той же выборке, затем выгрузка всех описаний. Индекс страны в план редакции не входит.

Почему. Заполнение всех карточек поднимает балл описаний и оставляет PDF на месте: критерий «файлом можно пользоваться» оно проигрывает. Новый API при пустых файлах отдаёт те же файлы быстрее. Ответы для индекса пишутся на уровне страны и региональный файл не чинят. Выборка с живыми файлами закрывает жалобу и укладывается в работу двух человек.

Последствия. Каталог ещё квартал будет выглядеть скромнее соседа с тысячами карточек. Выгрузка появится после файлов, и до неё федерация этот портал не заберёт.

Когда пересмотреть. Выборка файлов открывается программой, и национальный портал назвал дату сбора: тогда выгрузка становится первой задачей. Или писем о содержимом больше нет, а пустые поля остались у половины карточек: тогда очередь переходит к описаниям.

Как записать свою карточку — в главе «Как обосновать решение».

Если портала нет

Те же четыре уровня есть у страницы, репозитория и набора в общем хабе.

- **Данные** — сам файл: он открывается, колонки названы.
- **Описания** — README, лицензия, `datapackage.json` или разметка schema.org.
- **Сервис** — страница открывается без входа, адрес не меняется от правки заголовка.

- **Система** — запись в каталоге, который уже ищут: национальный портал, Zenodo, re3data. Поиск по GitHub эту запись не заменяет.

Порядок тот же: сначала файл, которым можно пользоваться, затем описание, затем постоянный адрес, затем запись уровнем выше. Минимум страницы — в главе о находимости без портала. Чем короткий проход отличается от аудита портала — в разделе «Если портала нет».

Как поставить задачу

В задаче пишут уровень, охват и исполнителя.

Годная формулировка выглядит так: «На выборке из 30 наборов разных издателей файл открывается в CSV или GeoPackage без ручной правки. Срок — месяц. Делает владелец данных, редактор проверяет».

Итог прохода — одна страница из главы о самоаудите. Очередь на ней уже разнесена: что правит редактор карточек и что требует смены выгрузки или адреса.

Источники

- T. E. Hardwicke и др., Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: an observational study, Royal Society Open Science, 2021 — 25 статей со значком открытых данных за 2014–2015 годы; без помощи авторов проверенные числа сошлись у 9, после пояснений авторов — у 15
- Пять звёзд открытых данных
- MQA, методика MQA
- Open Data Maturity, Global Data Barometer, ODIN
- CoreTrustSeal
- data.gouv.fr: руководство по баллу качества метаданных
- И. Бегтин, Что не так с порталом открытых данных Узбекистана?, февраль 2023 — размеры наборов data.gov.uz
- Метрики, методики и рейтинги, самоаудит
- Хорошие практики, федерация, как сделать публикацию находимой
- OpenNeuro, руководство пользователя — проверка BIDS при загрузке; DANDI, публикация наборов — публикация после исправления всех ошибок проверки; каталог схем schema.data.gouv.fr

Метрики качества

Метрика полезна, если по ней видно, что чинить. Для этого считают несколько долей. У каждой доли три записи: что стоит в знаменателе, какое значение считается заполненным и на какую дату посчитано. Число наборов в подвале сайта этих записей не имеет.

Из чего состоит доля

Знаменатель — наборы из названного списка: вся выгрузка, каждый N-й набор или страница API с указанным сдвигом. Долю считают от числа наборов в этом списке. Тридцать карточек разных издателей ловят типичную дыру. Как их взять и куда записать, — в главе о самоаудите.

«Заполнено» бывает трёх степеней. В отчёте пишут, какую считали. Иначе две доли с одним названием нельзя поставить рядом.

Степень	Что прошло	Что ещё может быть пустым по смыслу
Поле непустое	в лицензии что-то стоит	значение «не указана»: это тоже строка
Значение из списка	лицензия из известных, формат из справочника	период «1900–2099», описание, скопированное из названия
Значение правдиво	период похож на данные, ссылка открывает файл	это смотрят на нескольких карточках глазами

Первая степень даёт высокую долю там, где пустое поле заменили заглушкой. Вторая ловит заглушку, если она в списке запрещённых значений. Третью автоматическая проверка почти не закрывает: валидатор профиля и MQA смотрят, что поле есть. На data.gouv.fr встроенная оценка засчитывает лицензию «не указана», на data.gov.gr период-заглушка стоит почти у всех наборов выборки и уходит в выгрузку DCAT. Оба разбора — в главе о полноте метаданных.

Одна страница API — не выборка

Портал на CKAN отдаёт карточки страницами. Страница часто оказывается одной партией сбора, а не срезом каталога. 14 октября 2026 года API data.gov.gr показывал 22 765 наборов. За время проверок счётчик сдвигался на одну-две карточки. Команда ниже берёт 200 наборов, упорядоченных по дате создания карточки, и считает четыре числа: сколько организаций на странице, у скольких явную лицензию, у скольких хотя бы один файл из короткого списка форматов и у скольких карточка менялась после 14 октября 2025 года.

Явная лицензия здесь — вторая степень: поле `license_id` непустое и это не `notspecified` и не `none`. Код `other-open` («какая-то другая открытая лицензия») команда засчитывает. Более строгая проверка смотрит, есть ли у лицензии ссылка на текст. Короткий список форматов — CSV, JSON, XML, GeoJSON, GeoPackage, Parquet. Команда читает подпись формата, а не байты: ZIP и XLSX в список не входят, даже если внутри лежит таблица. Для геопортала список другой. Когда файл вообще пригоден программе — в главе о форматах.

```
PORTAL=https://data.gov.gr
START=0
CUTOFF=2025-10-14
curl -s "$PORTAL/api/3/action/package_search?
rows=200&start=$START&sort=metadata_created%20asc" \
  | jq --arg cut "$CUTOFF" '.result.results as $r | ($r | length) as $n | {
      datasets: $n,
      organizations: ([$r[].organization.name] | unique | length),
      explicit_license: ([$r[] | select((.license_id // "" | ascii_downcase) as $l
                          | $l != "" and $l != "notspecified" and $l != "none")] | length),
      open_format: ([$r[] | select(any(.resources[];
                     (.format // "" | ascii_upcase | gsub("\\."; "")) as $f
                     | ["CSV","JSON","XML","GEOJSON","GPKG","PARQUET"] | index($f)))] |
length),
      card_modified_since_cutoff: ([$r[] | select((.metadata_modified // "") > $cut)] |
length)
    }'
```

`CUTOFF` — дата год назад от дня проверки. Её подставляют свою. `START` меняют и запускают команду несколько раз.

Три запуска в тот день:

Сдвиг	Организаций	Явная лицензия	Открытый формат	Карточка менялась за год
0	75	67	18	200
8000	14	3	57	200
16000	1	0	200	200

Все числа — из 200. На второй странице 14 организаций, и 165 наборов из 200 пришли одним сбором `data-apdkritis`: карточки созданы 14 марта 2026 года. Третья страница — одна организация `keratsini-drapetsona`, все 200 карточек созданы 22 мая 2026 года, источник сбора называется `gis-keratsini-drapetsona-wms-only`. Поле формата при этом попало в короткий список у всех 200, а явной лицензии нет ни у одной. Доля «открытый формат 100%» описывает эту партию, а не портал.

Сортировка по идентификатору перемешивает издателей. Три страницы по 200, сдвиги 0, 8000 и 16000: организаций 70, 80 и 76, явная лицензия у 39, 41 и 36, открытый формат у 116, 110 и 115. Разброс меньше, но это всё ещё три страницы, а не перепись 22 765 наборов. Крупные издатели занимают на каждой странице заметную долю. Для отчёта берут каждый N-й набор, как в самоаудите, и пишут правило отбора.

В тот же день API open.canada.ca показывал 48 083 набора. Три страницы по идентификатору, по 200 карточек: явная лицензия у всех 200, открытый формат — у 88, 83 и 92. Лицензия на этих страницах уже не отличает каталог от самого себя. Смотреть стоит долю, которая ещё может провалиться.

Дата в команде тоже легко врёт. Поле `metadata_modified` за последний год заполнено у всех 200 карточек во всех трёх греческих срезах, включая партию одного муниципалитета. По этому полю портал выглядит полностью свежим. У набора `dedo` карточка в `metadata_created` датирована 17 ноября 2025 года, а у файла в поле `created` стоит 15 июня 2018 года. После

переноса портала дата рождения карточки — дата переноса. Свежесть данных смотрят по дате файла или по обещанной периодичности. Как портал сам считает срок следующего файла, на примере HDX, — в главе о хороших практиках.

При повторном запуске граница страницы может сдвинуться на карточку: в каталог продолжают добавлять наборы. В отчёт пишут дату, сдвиг и сортировку.

Что считать

Для городского и ведомственного портала хватает пяти чисел. Их пишут на страницу итога из главы о самоаудите.

Число	Как считать	Когда это уже очередь
Явная лицензия	доля наборов, где лицензия — ссылка на известный текст, а не пустое поле и не «не указана»	меньше половины выборки
Живой файл	доля наборов, где ссылка открыла файл	заметная доля ведёт на страницу сайта или не отвечает
Открытый формат	доля наборов с хотя бы одним файлом из записанного вами списка	список не записан, и долю нельзя повторить
Свежесть файла	доля наборов, у которых дата файла укладывается в обещанную частоту	дата есть только у карточки
Общая выгрузка	да или нет: все описания открываются без входа	выгрузки нет. Это не доля, а порог

Пороги «хорошо» и «плохо» задают под свою задачу. Лицензия меньше чем у половины выборки уже очередь на исправление, а не повод спорить о весах.

Рядом с этими пятью держат отдельный список, его в ту же долю не смешивают.

Что ещё	Что записывают	Чего число не видит
Полнота остальных полей	издатель, контакт, период, территория — каждое своей долей и своей степенью «заполнено»	скопированный текст, если его не искали отдельно
Совместимость	проходит ли выгрузка валидатор своего профиля	правду в полях, которые формально на месте
Использование	скачивания, вызовы API, ссылки из статей — вместе	качество файла. Большое число просмотров его не заменяет
Содержимое	доля коротких файлов и общий объём	это другой уровень качества. Порядок — в проверке по существу

Единица в знаменателе зависит от типа. Для таблиц это набор, для каталога показателей — ряд, для геопортала — слой, для микроданных — исследование. Что считать по файлам у каждого типа, собрано в той же главе о самоаудите.

Чем считать

Поля и ссылки считает программа. Совпадение описания с соседней карточкой и правдоподобие периода — глаза, на тех же тридцати карточках.

Адреса ниже отвечали 14 октября 2026 года.

Что проверяет	Инструмент
Выгрузка соответствует DCAT-AP	валидатор DCAT-AP на платформе тестирования Еврокомиссии
Описания слоёв и сервисы соответствуют INSPIRE	INSPIRE Reference Validator: центральный экземпляр Еврокомиссия закрыла 1 апреля 2026 года, код и наборы тестов открыты, валидатор разворачивают у себя на ETF
Разметка schema.org на странице набора читается	Schema Markup Validator и проверка расширенных результатов Google
Запись научного репозитория похожа на FAIR	F-UJI — по идентификатору или адресу записи. Содержимое файла программа не читает
Карточки, которые забрал европейский портал	MQA на data.europa.eu. Каталог, который туда ничего не отдаёт, балла не получает
Ссылки на файлы открываются	проверка адреса. Код ответа 200 ещё не значит, что по ссылке файл

Проверка ссылок отличает четыре исхода: файл открылся, сервер отказал, ответа нет, открылась страница сайта. 14 октября на дюжине ссылок с одной страницы open.canada.ca девять ответили успехом, одна — отказом 403, две не ответили за 15 секунд. На дюжине ссылок с data.gov.gr все двенадцать ответили успехом, и одна вела на HTML-раздел сайта. Это проверка приёма, а не доля по каталогу. Адрес на CKAN бывает путём от корня сайта, без `https://`. Проверка, которая требует полное имя сайта, посчитает живую ссылку пустой. Мягкая поломка, когда ссылка открывается и ведёт не на файл, разобрана у термина «битая ссылка» в глоссарии.

Как устроена формула MQA

MQA — образец чужой формулы, а не обязательный балл своего портала. Как читать балл каталога и почему он покрывает только посчитанные описания, — в главе о методиках. Здесь — из чего складывается одно слагаемое. Текст ниже сверен с разделом методики в отчёте каталога от 12 октября 2026 года. Страница методики на сайте датирована 6 октября 2026 года.

Проверки подписаны четырьмя словами FAIR: находимость, доступность, совместимость, повторное использование. В прежней методике было пятое измерение, contextuality. В текущей его нет. Вес берётся из обязательности поля в DCAT-AP, а не из отдельного экзамена на FAIR: обязательное поле — 1, рекомендуемое — 0,5, необязательное — 0,25. У набора, способа получения и сервиса данных потолок один и тот же, 7,5. Балл каталога — среднее на той же шкале. Полосы на шкале: до 2,5 — достаточно, от 2,5 до 5 — хорошо, от 5 до 7,5 — отлично.

Большинство проверок устроены одинаково: свойство заполнено или нет. Заглушка такую проверку проходит. Живая ссылка — исключение, и оно слабее, чем кажется. Наличие адреса доступа с правильным видом URL весит 1. Проверка, что адрес отвечает, весит 0,25, даже когда само поле обязательное. В зачёт идёт только ответ с кодом 2xx на запрос заголовка. Файл при этом не скачивают. Перенаправление, которым порталы часто отдают файл, в зачёт не идёт.

Соответствие профилю SHACL в балл не входит. Его смотрит отдельный валидатор из таблицы выше. Содержимое файла, юридическую силу лицензии и правду в полях методика не проверяет.

Свою формулу можно собрать короче: вес по обязательности поля и проверка, что ссылка открыла файл. Повторять шкалу 7,5 незачем. Прежде чем ставить чужой балл в отчёт, формулу читают. Балл карточки на data.gouv.fr считает поле лицензии заполненным и тогда, когда в нём «не указана». Как устроен тот балл — в главе о хороших практиках.

Что легко подделать

Число наборов. Его поднимают, режа один ряд на файлы. Каталог из 50 честных наборов полезнее каталога из 5 000 пустых карточек. Если число всё же нужно для отчёта, рядом ставят два других: долю коротких файлов и общий объём. Как их считать — в проверке по существу. Где проходит граница между нарезкой и отдельным объектом — в антипаттерне «Дробление ради счётчика».

Насколько счётчик расходится с объёмом, видно по одному снимку. 2 февраля 2022 года полная выгрузка data.gov.ru дала 26 490 наборов и 29 ГБ, в основном без сжатия. 6,8 ГБ из них — один слепок адресного реестра, не обновлявшийся с октября 2019 года. 3839 наборов, 14,5%, весили меньше килобайта. Сжатый архив того же снимка — 14 ГБ: в других главах это те же данные, уже упакованные. 14 февраля 2022 года у HDX при сопоставимом числе наборов, около 31 тысячи, скачанная часть — 6600 наборов, 21% — занимала 205 ГБ. Оценка всего каталога по этой части — 950–1000 ГБ, это оценка незаконченной загрузки. У data.gov.uk в той же загрузке одни CSV без сжатия занимали 222 ГБ, и это не весь портал. Что осталось на data.gov.ru после открытия в 2025 году — в главе о самоаудите.

Поэтому показатель «опубликовать N наборов» без перечня наборов ничего не меряет: такую цифру собирают за день из коротких таблиц. Осмысленная цель — список конкретных наборов, например наборов высокой ценности, и доля опубликованных из него.

Число просмотров главной страницы. Оно не видит тех, кто забирает выгрузку. Если основной вход — API, считают и его. Как не превратить статистику в годовую цифру ради цифры — в главе о прочих возможностях.

Балл за поле, которое формально на месте. Валидатор и MQA проверяют наличие. Они не проверяют, что в описании не скопирован текст соседнего набора и что период не заглушка. К автоматической оценке добавляют просмотр нескольких карточек.

Дата, которую двигает перенос или мониторинг. Правка карточки и повторная загрузка каталога обновляют `metadata_modified`, не обновляя строки. Раздел мониторинга может ставить «обновлено» набору из одной старой цифры. Так было на data.gov.uz. Разбор — в самоаудите, в разделе о том, где проход обманывает.

Признаки жизни

Отдельная задача — понять, жив ли портал вообще. Для чужого каталога, на который собираетесь опираться, и для своего, если его давно никто не открывал, хватает пяти проверок. Их делают до полной выгрузки файлов.

- дата последнего **нового** набора, и это не дата переноса портала: у старого файла дата в карточке может быть вчерашней;
- дата последней новости или записи в журнале изменений;

- есть ли ряды с обещанной частотой и совпадает ли дата файла с этой частотой;
- доля ссылок, которые открывают файл, а не страницу;
- не пропали ли разделы, которые были раньше: план публикации, статистика использования, запросы данных.

Когда весной 2022 года с data.gov.ru пропала страница статистики посещений, а план публикации data.mos.ru удалили ещё в 2020 году, это было заметнее любого рейтинга. Для государственных порталов 2020–2022 годов был простой исторический тест: публиковали ли они данные о COVID-19, пока те были нужны. Таиландский data.go.th обновлял такую статистику каждый день, российский data.gov.ru — нет. Сегодня тот же вопрос задают ряду, который и сейчас кому-то нужен по расписанию: бюджет, курс, еженедельный оперативный набор. Портал, на котором такой ряд обновляется в обещанный день, жив независимо от счётчика наборов.

Источники

- Методика MQA — страница от 6 октября 2026 года; раздел методики в отчёте каталога от 12 октября 2026 года: четыре измерения FAIR, веса 1 / 0,5 / 0,25, шкала 0–7,5, проверка ссылки кодом 2xx, SHACL в балл не входит
- Валидатор DCAT-AP, F-UJI — адреса открывались 14 октября 2026 года; INSPIRE Reference Validator — объявление JRC о закрытии центрального экземпляра, 1 апреля 2026 года, ETF и исполняемые наборы тестов для своей установки
- Schema Markup Validator, разметка Dataset в документации Google
- Action API CKAN 2.12
- API data.gov.gr и open.canada.ca, 14 октября 2026 года: срезы по 200 наборов, сортировка по дате создания карточки и по идентификатору
- И. Бегтин, Telegram: полная выгрузка data.gov.ru, 02.02.2022; сравнение объёмов порталов, 14.02.2022; показатель «500 наборов», 21.03.2025; data.mos.ru по инерции, 17.01.2022; статус порталов весной 2022 года, 05.05.2022; данные о COVID-19 как признак живого портала, 15.12.2021, и data.go.th, 17.07.2022
- Полнота метаданных, самоаудит, методики и рейтинги, антипаттерны

Методики и рейтинги

Балл отвечает на один вопрос. У Франции в отчёте Open Data Maturity за 2025 год стоит 100%. У того же национального портала в оценке карточек data.europa.eu 11 октября 2026 года было 4,27 из 7,5, и в среднее вошли 211 описаний, хотя в поиске европейского портала этот каталог числится на 76 548 наборов. Оба числа можно цитировать. Подставлять одно вместо другого нельзя: самоаудит одного сайта тоже отвечает на свой вопрос, он разобран в соседней главе.

Сначала таблица. Год — выпуск, на который можно ссылаться сейчас.

Методика	Единица	Что меряет	Что остаётся за кадром
Пять звёзд (2010)	один файл	лестницу формата: от лицензии к связанным данным	портал, свежесть, живые ссылки
FAIR (2016) и F-UJI	одна запись	поля, идентификаторы и условия доступа	содержимое файла и портал открытых данных
MQA	карточки, которые забрал data.europa.eu	заполненность DCAT-AP и ответ ссылок	наборы, которых в расчёте нет, и строки внутри файла
Балл карточки на своём портале	одна карточка	поля, которые портал сам решил проверять	чужие каталоги и содержимое файла
CoreTrustSeal	репозиторий как архив	процедуры, сохранность, права	каждый файл и городской бюджет
Open Data Maturity (2025)	страна Европы	политику, портал, контроль качества и влияние по анкете	конкретный CSV
OURdata (2025)	центральное правительство страны ОЭСР	доступность данных, доступ и поддержку повторного использования	влияние и городской сайт
Global Data Barometer, 2-е изд. (2025)	страна Африки или Латинской Америки	данные для подотчётности, управление и навыки	Европу в этом выпуске
ODIN 2024/25	сайт статистической службы	полноту и открытость официальной статистики	портал открытых данных, если служба на него не ссылается
Балл доверия реестра	карточка каталога	грубый фильтр тысяч каталогов снаружи	обновляются ли наборы и открывается ли файл

Как читать чужой балл

Прежде чем ставить число в отчёт, отвечают на четыре вопроса.

1. **Единица.** Файл, запись, карточки которые забрал верхний портал, политика страны или сайт статистической службы.
2. **Кто заполнял.** Программа, сама страна или эксперт по открытым источникам.
3. **Какой выпуск.** Если между выпусками сменились вопросы или веса, прошлогодний балл с нынешним не складывают.
4. **Сколько записей внутри балла.** Среднее по двумстам карточкам и среднее по всему каталогу — разные утверждения. Рядом с баллом пишут это число.

Дальше — чем каждая семья отличается. Живой разбор трёх каталогов в европейской федерации — в разделе про MQA.

Один файл

Пять звёзд открытых данных (Тим Бернерс-Ли, 2010) — лестница для одного набора.

Звезда	Что уже есть
1	Файл лежит в вебе под открытой лицензией. Формат любой, хоть PDF
2	Это структура, которую читает программа: таблица, а не снимок таблицы
3	Формат открытый: CSV, JSON или XML, а не файл одной платной программы
4	У сущностей внутри данных есть адреса, на которые можно сослаться
5	Эти адреса ведут на чужие данные

Одна образцовая таблица набирает пять звёзд при мёртвом каталоге. Лестница годится для разговора о формате. Аудитом портала она не работает.

В анкете Open Data Maturity страну спрашивают, пользуется ли она этой лестницей или похожей моделью. Балл получает ответ национальной команды, а не подсчёт звёзд по файлам.

Одна запись

Принципы FAIR (2016) говорят, что научные данные должны находиться, быть доступны, стыковаться с другими и использоваться повторно. Это требования к записи и к условиям доступа. FAIR и «открыто» — разные свойства: набор с закрытыми файлами тоже может им соответствовать, если описание открыто, идентификатор постоянный, а порядок доступа понятен. Для городского CSV принципы задают словарь. Лицензию и выгрузку DCAT они не заменяют.

F-UJI ставит автоматический балл одной записи по идентификатору или адресу страницы. Программа смотрит поля и идентификаторы. Содержимое файла она не читает. Как прогнать запись и чем ещё считать качество — в главе о метриках. Где FAIR обязателен репозиторию — в главе о научных репозиториях.

Карточки, которые забрал верхний портал

MQA на data.europa.eu оценивает описания, которые порталы отдают в европейский каталог. Каталог, который туда ничего не отдаёт, балла не получает, каким бы аккуратным он ни был.

Шкала каталога — от 0 до 7,5. Вес поля взят из обязательности в DCAT-AP: обязательное — 1, рекомендуемое — 0,5, необязательное — 0,25. Проверяют три части описания: набор, способ получения и сервис данных. Балл каталога — среднее по тем наборам, которые оценка посчитала, плюс отдельные сервисы данных, которые лежат в каталоге сами по себе. Устаревший способ получения в среднее не входит. Ссылка считается доступной, если запрос заголовка получает код 2xx. Файл при этом не скачивают, а перенаправление в зачёт не идёт. Содержимое файла, юридическую силу лицензии и правду в описании методика не проверяет.

Проверки по-прежнему подписаны четырьмя словами FAIR: находимость, доступность, совместимость, повторное использование. Вес при этом берётся из обязательности поля, а не из отдельного экзамена на FAIR. Страница методики датирована 6 октября 2026 года. Свою формулу можно собрать по тому же принципу: вес по обязательности и проверка живых ссылок. Подробнее — в главе о метриках.

Балл читают вместе с числом описаний в расчёте. 14 октября 2026 года в кэше MQA балл был у 200 каталогов из 215. Даты расчёта в этом кэше — 23–24 сентября и 11 октября. Среднее баллов каталогов — 4,75 из 7,5, медиана — 4,91. Это среднее каталогов, а не качество всех европейских наборов.

Три каталога на снимке MQA от 11 октября 2026 года. Число наборов в поиске data.europa.eu снято 14 октября.

Каталог	Балл	Описаний в расчёте	Наборов в поиске
data.gov.uk	2,66	9 505	59 041
data.gouv.fr	4,27	211	76 548
dati.gov.it	6,69	61 693	61 748

У британских карточек, которые вошли в балл, прямой ссылки на файл не было ни у одного способа получения. Лицензии из европейского справочника — тоже. Поле лицензии заполнено у 25,9% способов получения. Адрес страницы есть у 99,9%: это обязательное поле и оно тяжёлое. Сам адрес ответил у 35,4%. Тема набора заполнена у 3,5%, контакт — у 0,9%, период и территория — ни у кого. Отсюда 2,66 при почти десяти тысячах описаний в расчёте. В поиске европейского портала у того же каталога 59 041 набор, так что и этот балл покрывает долю сбора. Великобритания в Open Data Maturity не участвует: это другая шкала и другой круг стран.

Итальянский балл 6,69 стоит на почти полном сборе: в расчёт вошли 61 693 описания из 61 748 в поиске. Лицензия и прямая ссылка есть у каждого способа получения. Это верхняя часть шкалы карточек. Про строки внутри файлов она по-прежнему молчит.

Французские 4,27 посчитаны по 211 описаниям. Рядом в поиске — 76 548 наборов того же каталога, «открытой платформы французских публичных данных», то есть data.gouv.fr. Цитировать 4,27 как оценку этих 76 тысяч нельзя. Сначала смотрят, сколько записей вошло в среднее.

Команды для своего каталога, если он уже есть в data.europa.eu. Идентификатор британского — `data-gov-uk`, французского — `plateforme-ouverte-des-donnees-publiques-francaises`, итальянского — `dati-gov-it`.

```
curl -s "https://data.europa.eu/api/mqa/cache/catalogues/data-gov-uk" \
  | jq '{score: .result.score, assessed: .result.counts.value}'
curl -s "https://data.europa.eu/api/hub/search/catalogues/data-gov-uk" \
  | jq '{listed: .result.count, home: .result.homepage}'
```

Первая показывает балл и число описаний в расчёте. Вторая — сколько наборов европейский портал держит в поиске. Если второе число намного больше первого, балл описывает кусок сбора.

В отчёте Open Data Maturity за 2025 год пять метрик MQA стоят рядом с анкетой как пилот, по одному каталогу на страну. В четверть балла страны они не входят. Туда же не входят автоматические проверки главной страницы портала: мобильная вёрстка, скорость, безопасность и доступность. Измерение «качество» в том отчёте — ответы о том, следит ли страна за метаданными. Об этом следующий раздел.

Балл на своём портале

Портал может показывать издателю балл карточки до того, как её заберёт кто-то сверху. На data.gouv.fr в него входят длина описания, описание файлов, частота и её соблюдение, открытая лицензия, открытый формат, территория и период. Балл с 2022 года экспериментальный. Как он устроен и что 14 октября 2026 года было закрыто у карточки самого каталога — в главе о хороших практиках. Это третье число рядом с анкетой страны и с MQA: его видит издатель на своём сайте.

Архив

CoreTrustSeal сертифицирует репозиторий как место, которому можно доверить данные надолго: процедуры, сохранность, права. Городскому порталу бюджетов сертификат архива не нужен. Сколько репозиториев его держат и какие требования входят в анкету — в главе о научных репозиториях. В реестре re3data отметка о сертификате стоит рядом с карточкой. Это сигнал для того, кто выбирает, куда положить данные на годы.

Страна

Индексы сравнивают политику и наличие данных в стране. Балл получает страна, а не городской сайт. Общее ограничение одно: высокий балл оставляет открытым вопрос, открывается ли конкретный файл.

Open Data Maturity — ежегодная оценка с 2015 года, одиннадцатый выпуск вышел в декабре 2025-го. Участие добровольное. С 2018 года балл складывается из четырёх равных частей: политика, портал, качество и влияние. В 2025 году вопросы не меняли; последняя перестройка анкеты была в 2024 году.

Это опрос национальных представителей. Большинство вопросов — выбор из готовых ответов плюс ссылка или объяснение. Команда data.europa.eu проверяет, что объяснение относится к вопросу. С 2024 года анкету заранее заполняют прошлогодними ответами, и подробно разбирают то, что страна изменила, и новые вопросы. Повторный «да» заново по файлам не обходят.

В 2025 году участвовали 36 стран: 27 стран ЕС, три страны ЕАСТ (Исландия, Норвегия, Швейцария) и шесть кандидатов (Албания, Босния и Герцеговина, Черногория, Северная Македония, Сербия, Украина). Средний балл всех участников — 81%, на один пункт выше 2024 года. Средний балл ЕС-27 — 86%, на три пункта выше. Внутри ЕС-27 выше всего политика (93%) и портал (85%), затем качество (83%) и влияние (82%). Выше 80% — 21 страна из 36. Франция набрала 100% по темам анкеты, Литва и Польша — 98%. Среди стран ЕАСТ выше

всех Норвегия (92%), среди кандидатов — Украина (97%). Сильнее всего выросли Албания (на 23 пункта), Мальта (на 19) и Германия (на 11). У одной страны балл снизился. У Черногории и Северной Македонии сравнения с 2024 годом нет: они вошли в оценку впервые.

Скачок на 23 пункта за год — про ответы и примеры, которые страна смогла показать. Измерение «качество» спрашивает, есть ли порядок обновления метаданных, следят ли за лицензиями и за DCAT-AP, известна ли доля данных, которая проходит выбранную модель вроде пяти звёзд. Оно не равно баллу MQA.

Ответ «к нам эта норма ЕС не относится» по вопросам о законодательстве ЕС даёт полный балл. Так устроены сравнения со странами ЕАСТ и кандидатами: их высокий процент включает вопросы, которые для них сняты.

OURdata ОЭСР — открытость государственных данных в странах ОЭСР. Выпуск 2025 года опубликован 16 февраля 2026 года, на октябрь 2026 года он последний. Данные собраны опросом центральных ведомств в 2025 году и описывают политику с 1 января 2023 по 31 декабря 2024 года. Три равные части: доступность данных, удобство доступа и поддержка повторного использования со стороны правительства. Шкала от 0 до 1. Влияние открытых данных индекс не меряет. Средний балл стран ОЭСР вырос с 0,48 в выпуске 2023 года до 0,53. Выше всех Франция (0,96) и Корея (0,95). Данных 2025 года нет по Дании, Венгрии и США. Единица — центральное правительство, городской сайт в балл не входит.

Global Data Barometer смотрит данные, которые нужны для подотчётности: закупки, компании, земля, политическая добросовестность, общественные финансы, плюс управление данными и навыки. Второе издание вышло 29 мая 2025 года. В нём 43 страны Африки и Латинской Америки с Карибским бассейном, 27 основных показателей, период с 31 августа 2022 года по 1 сентября 2024 года. Первое издание, май 2022 года, охватывало 109 стран и период с мая 2019 по май 2021 года. Баллы двух изданий не складывают: число показателей, веса и правила подсчёта изменились. У европейской страны во втором издании балла нет. Цифра из первого издания — снимок 2019–2021 годов.

ODIN (Open Data Watch) оценивает официальную статистику. Выпуск 2024/25 — седьмой, в нём 197 стран, страница рейтинга обновлена 21 июля 2025 года. Смотрят сайт национальной статистической службы и официальные страницы, на которые она ссылается. Полнота — есть ли данные по 22 темам, с разрезами, за ряд лет и по территориям. Открытость — машиночитаемый и открытый формат, описание, способ скачать выбранное или всё сразу, открытая лицензия. Средний балл стран вырос на 12,5% к 2022 году. Средняя открытость — 59 из 100: данные, которых нет, на открытость не оценивают, а у тех, что есть, формат и лицензия часто ещё закрыты.

Узбекистан в этом выпуске набрал 83 балла: полнота 78, открытость 87. Это сайт статистической службы. Портал открытых данных data.gov.uz в феврале 2023 года на 96% состоял из коротких файлов — разбор в главе о самоаудите. Два сайта, два балла.

Старые индексы полезны как история метода. **Open Data Barometer** Всемирного фонда веба: четвёртое, последнее полное издание вышло в мае 2017 года (период июль 2015 — июнь 2016), выпуск для 30 правительств — в сентябре 2018 года (период июль 2016 — сентябрь 2017). Глобальных выпусков после этого нет. Global Data Barometer вырос из этой работы и

ведёт свой счёт. **Global Open Data Index** Open Knowledge: последнее издание — 2016/17, опубликовано в июне 2017 года, 94 территории. Выпуска 2018 года не было, сайт архивный. Ссылаться на эти баллы как на текущие нельзя.

Грубая оценка в реестре каталогов

Реестр dataportals-registry считает необязательный балл доверия от 0 до 100. Это не сертификат. Балл складывается из пяти частей: тип владельца (наука выше бизнеса, центральное правительство выше местного), тип каталога (репозиторий чуть выше портала, агрегатору и маркетплейсу балл снижают), лицензия, карточка в re3data и сертификат, живы ли сайт и API. Пустое описание лицензии балл снижает. Итог обрезают до диапазона от 0 до 100.

Тип владельца весит больше остальных: у науки 40 баллов, у сообщества — 5. Тип «федеральное правительство» в таблице баллов не назван, поэтому такой каталог получает те же 10 баллов, что и пропуск в поле владельца. Низкий балл федеральной площадки может быть дырой формулы, а не приговором каталогу.

Балл поощряет заполнить карточку каталога. Он не видит, обновляются ли наборы и открывается ли файл. Для самоаудита своего портала он слишком грубый. Для первого отбора тысяч каталогов снаружи — годный фильтр. Формула — в документации реестра, ссылка ниже.

Что из этого брать себе

Ситуация	Что открывать
Чините городской или ведомственный портал	самоаудит: файлы, карточки, выгрузка. Индекс страны к одному CSV не прикладывают
Вас уже собирает data.europa.eu	MQA, и рядом — число описаний в расчёте. Если оно далеко от числа наборов в поиске, балл описывает кусок
Ведёте научный архив	FAIR для записи и, когда архиву доверяют данные на годы, CoreTrustSeal
Отвечаете за открытые данные страны в Европе	Open Data Maturity — чтобы сравнить политику и портал с соседями. Качество файлов всё равно смотрят у себя
Сравниваете официальную статистику стран	ODIN. Это сайт статистической службы, не обязательно национальный портал
Выбираете чужой каталог из тысяч	балл доверия реестра как первый фильтр, затем признаки жизни из главы о метриках

Городской портал не поднимает индекс страны в одиночку. Индекс страны не открывает городской файл. Задачу ставят на той единице, где сломано.

Источники

- Пять звёзд, Тим Бернерс-Ли, 2010; принципы FAIR; M. D. Wilkinson и др., The FAIR Guiding Principles, Scientific Data, 2016; F-UJI
- CoreTrustSeal; число сертифицированных репозиториев и требования 2026–2028 годов — в главе о научных репозиториях

- MQA, методика по состоянию на 6 октября 2026 года; кэш каталогов и поиск data.europa.eu — 14 октября 2026 года, даты расчёта баллов в кэше — 23–24 сентября и 11 октября. Каталоги: `data-gov-uk` (2,66 из 7,5 по 9 505 описаниям, в поиске 59 041), `plateforme-ouverte-des-donnees-publiques-francaises` (4,27 по 211 описаниям, в поиске 76 548), `dati-gov-it` (6,69 по 61 693 описаниям, в поиске 61 748). Среднее баллов каталогов — 4,75 из 7,5, медиана 4,91, балл есть у 200 из 215
- Open Data Maturity, отчёт 2025 года — 36 стран, ЕС-27 86%, все участники 81%, Франция 100%; методика 2025 года — четыре измерения по 25%, опрос с предзаполнением, пилотные метрики MQA в балл не входят
- ОЭСР, Digital Government Index and OURdata Index: 2025 Results and Key Findings, 16 февраля 2026 года — средний балл стран ОЭСР 0,53, данных нет по Дании, Венгрии и США; предыдущий выпуск — OURdata Index 2023, 22 декабря 2023 года, средний балл 0,48
- Global Data Barometer, второе издание, 29 мая 2025 года — 43 страны; первое издание, май 2022 года — 109 стран; сравнение изданий
- ODIN 2024/25 — 197 стран, рост среднего балла на 12,5% к 2022 году, средняя открытость 59; рейтинг, обновлён 21 июля 2025 года — Узбекистан 83 (полнота 78, открытость 87)
- Open Data Barometer — 4-е издание, май 2017 года; выпуск лидеров, сентябрь 2018 года. Global Open Data Index — архив, последнее издание 2016/17, июнь 2017 года, 94 территории
- Оценка доверия в реестре
- Метрики, самоаудит, уровни качества, хорошие практики

Самоаудит каталога

Самоаудит — один проход по каталогу, после которого ясно, что чинить в ближайший месяц. Итог умещается на страницу: когда смотрели и по какому списку, несколько долей, пять карточек с типичными ошибками и две очереди. Одна — редактору карточек, вторая — тому, кто отвечает за платформу.

Рейтинг страны и сертификат архива отвечают на другие вопросы. Их место разобрано в главах об уровнях качества и о методиках. Здесь — порядок для своего портала и для чужого, на который собираетесь опираться.

Полный проход делают раз в полгода и в день смены платформы или адреса: старые ссылки ломаются сразу. Раз в месяц достаточно открыть десяток известных ссылок. Квартальные и годовые проверки платформы — в главе о жизненном цикле. Смотрят двое. Строки и карточки — тот, кто отвечает за данные. Выгрузку, адреса и версию — тот, кто отвечает за платформу.

Два масштаба

Выборка из 30 наборов ловит типичную ошибку карточки: нет лицензии, период-заглушка, файл не открывается. Долю коротких файлов, общий объём и год публикации считают по выгрузке файлов, лучше по всем. Ниже два портала, где паспорта выглядели живыми, а подсчёт файлов показал справки.

Масштаб	Что отвечает	Когда
30 карточек разных издателей	типичная ошибка описания и битая ссылка	каждый полный проход
Все файлы или каждый N-й, несколько сотен	распределение по размеру, общий объём, год файла, пропуски по главным публикаторам	когда счётчик наборов стал главным числом отчёта, при приёмке портала и после резкого роста этого счётчика

Долю из тридцати записывают как долю выборки: «в 30 наборах лицензии нет у 20». Если одна и та же дыра в большинстве тридцати, чинят правило заполнения, а не эти тридцать карточек.

Как взять тридцать

Берут из общей выгрузки, не с первой страницы сайта и не из блока «недавно обновлено». Выдача начинается с того, что редакция только что поправила. Удобный способ — упорядочить записи по идентификатору и взять каждый N-й, где N — число наборов, делённое на 30. У каталога из 6000 записей это каждый двухсотый.

Издатели должны различаться. Если девять десятых каталога — одно ведомство, тридцать карточек расскажут про него. Тогда к каждому N-му добавляют несколько наборов из длинного хвоста и так и пишут в поле «охват».

Выгрузки нет — те же тридцать набирают, листая общий список на сайте. В итог первой строкой ставят «общей выгрузки нет». Остальные метрики без неё всё равно считают, но каталог для федерации уже не годится.

Чем считать лицензию, дату, период и открытый формат и какими валидаторами прогонять профиль — в главе о метриках. Ошибки профиля (DCAT-AP, INSPIRE, DataCite) держат отдельным списком от пустых полей: валидатор проверяет, что поле есть, и молчит, если в соседних карточках один и тот же текст.

Шаги

1. Найдите каталог в профильном реестре, если он там есть, и проверьте адрес и тип. Ошибочная карточка путает тех, кто ищет портал. Число наборов на сайте и число в выгрузке или API записывают оба: они часто расходятся.
2. Откройте выгрузку всех описаний и страницу одного набора без входа. Регистрация, капча или проверка «я не робот» означают, что машинного доступа нет. Команды для карты сайта, `data.json`, API, OAI-PMH и CSW — в чек-листе живого портала.
3. Возьмите 30 наборов, как выше, и посчитайте по ним лицензию, дату, период и то, открывается ли файл. Сверьте дату в карточке с датой внутри файла. Поле вроде `metadata_modified` у CKAN — дата правки карточки.
4. Прогоните выгрузку через валидатор своего профиля. Список инструментов — в главе о метриках.
5. Если решаете вопрос объёма, посчитайте файлы, как в следующем разделе.
6. Выпишите пять–десять главных публикаторов темы или страны и отметьте, есть ли их данные. Пустая клетка напротив статистической службы или казначейства важнее сотни мелких наборов.
7. Сверьте результат с чек-листом требований. Пункт, который вы сами пометили обязательным и не выполнили, идёт в начало очереди.
8. Отложите пять карточек разных издателей со ссылками. Пять ошибок одного ведомства говорят о нём, а не о каталоге.

Чужой каталог проходят так же. Очередь тогда короче: брать его, идти в первоисточник или отложить. До выгрузки файлов имеет смысл пройти признаки жизни: мёртвый портал целиком не качают.

Проверка по существу

Четыре вопроса одни и те же у любого каталога. Меняется единица, которую считают.

Тип	Что считать	Куда смотреть команды
Таблицы, портал открытых данных	строки, сжатый объём, год файла	этот раздел
Каталог показателей	длина ряда и есть ли выгрузка всей базы	каталоги показателей
Геопортал	слой отдаёт объекты, а не только карточку	геопорталы
Научный репозиторий	за страницей записи есть файл	научные репозитории

Тип	Что считать	Куда смотреть команды
Микроданные	у исследования заполнены переменные и условия доступа	микроданные
ML-каталог	файл открывается без входа, лицензия на обучение названа	ML-каталоги

Для таблиц вопросы такие.

1. **Размер.** Разложите файлы по числу строк данных: пусто или одна строка, до 100, от 100 до 10 000, больше. Если большинство в первой группе, каталог состоит из коротких справок или один ряд разрезан на куски. Где проходит граница, — в антипаттерне «Дробление ради счётчика».
2. **Объём.** Сложите размер файлов и подпишите, сжаты они или нет. Сравнивают одинаковое с одинаковым и с одним известным источником той же темы: статистической базой, реестром ведомства, городским порталом. Каталог страны, который меньше одной её информационной системы, тему не покрывает.
3. **Возраст.** Год берут из файла или из поля первой публикации набора, а не из даты правки карточки. Новый портал, в котором почти всё впервые опубликовано много лет назад, перенёс старое.
4. **Охват.** Таблица публикаторов из шага 6. Её заполняют по именам организаций в выгрузке, а не по памяти.

Для папки с CSV первый подсчёт делается так. `wc` считает и строку заголовка, поэтому файл из одной строки данных даёт 2. Порог «не больше 2» как раз ловит пустые и однострочные файлы. На порогах 100 и 10 000 одна строка заголовка группу не меняет.

```
for f in data/*.csv; do wc -l < "$f"; done \
  | awk '{ n = $1
           if (n <= 2) a++
           else if (n <= 100) b++
           else if (n <= 10000) c++
           else d++ }
         END { print "пусто или 1 строка:", a+0,
                     " до 100:", b+0,
                     " до 10 000:", c+0,
                     " больше:", d+0 }'
```

`+0` нужен, чтобы пустая группа напечаталась нулём. Файл с расширением CSV, который открывается как страница сайта, в этот подсчёт не берут: это битая ссылка. Общий размер папки, как она лежит на диске:

```
du -ch data | tail -n 1
```

Если публикуете ZIP или gzip, в итог пишите размер архива, а не распакованной папки. Для XLSX и JSON строку считают программой, которой файл открывают. Пороги те же.

data.gov.uz, февраль 2023 и июль 2025

В феврале 2023 года на data.gov.uz было 6623 набора — больше, чем на большинстве порталов региона. Форматы машиночитаемые, у полей есть описания, раздел мониторинга помечал обновлёнными 6120 наборов. Выгрузка в CSV дала другое: 2823 набора (42,6%) состоят из одной строки, 6381 (96%) короче 200 строк, весь CSV занимает 40 МБ. Самый большой файл — 60 984 строки и 4,4 МБ в CSV. Набор «Получено яиц, всего по Хорезмской области» — одна цифра за 2020 год, в мониторинге он числился обновлённым 26 декабря 2022 года.

Рядом, в те же недели, у data.gov.kg было 646 наборов и около 570 МБ файлов, в основном Excel. Наборов меньше на порядок, объём больше. Форматы разные, поэтому это сравнение порядка, а не точных байтов.

В июле 2025 года повторный подсчёт, в заметке адрес указан как data.egov.uz: наборов 10 412, из одной строки — 5207, половина, больше 1000 записей — у 114, чуть более 1%. Выгрузка в JSON без сжатия — 426 МБ; для 2023 года в той же заметке названы 284 МБ JSON. Рядом с 40 МБ CSV эту цифру не ставят: формат другой. Число наборов выросло, правило нарезки осталось.

data.gov.ru, июль 2025

15 июля 2025 года портал открыли снова. Предыдущую версию закрыли в 2023 году «для переноса на новую платформу». Через API было 5826 наборов, на страницах сайта — меньше. Сжатый ZIP — около 100 МБ, без сжатия — около 1 ГБ. Из этих 5826 наборов 4694 (80%) короче 100 строк, 586 (10%) — от 100 до 500, 350 (6%) — от 500 до 10 000, 62 (около 1%) — от 10 000 строк. Около 3% скачать не удалось. По полю первой публикации 84% наборов относятся к 2019 году или раньше. Отметка «обновлён в 2025 году» саму строку в файле не меняет: в заметке такие наборы ещё предлагалось открывать руками.

Снимок 2 февраля 2022 года — 26 490 наборов и 29 ГБ, в основном без сжатия. Готовый архив того снимка — 14 ГБ; в разборе 2025 года его округлили как «больше 24 тысяч наборов». На новый портал он не вернулся. Для масштаба, из того же разбора июля 2025 года: выгрузка ЕМИСС — 6905 показателей и 4,36 ГБ сжатого CSV; снимок data.mos.ru в ноябре 2022 года — 988 наборов и 3 ГБ в сжатом виде. Данных Росстата и финансовых ведомств на новом портале не нашлось.

Оба разбора считали все файлы. Выборка описаний на таком портале этот провал не показывает.

Что писать в итоге

Одна страница. Её можно отдать редактору без пересказа. Поле, которое не смотрели, заполняют словами «не считали», а не прочерком: прочерк потом читают как «всё в порядке».

Поле	Что записать
Дата и охват	когда смотрели; все наборы или выборка, из какого списка, сжаты ли файлы
Лицензия	доля выборки, где лицензия есть ссылкой
Живая ссылка	доля выборки, где файл открылся
Общая выгрузка	есть или нет; открывается ли без входа; совпало ли число с сайтом

Поле	Что записать
Содержимое	доля пустых и однострочных, общий объём с пометкой «сжато» или «как лежит»
Пять карточек	типичные ошибки со ссылками
Очередь на месяц	что правит редактор карточек и владельцы данных
Очередь на платформу	что требует смены расширения, выгрузки или адреса

Так выглядит заполненная страница по проходу февраля 2023 года. Лицензию и долю битых ссылок в том разборе не считали — в таблице это видно.

Поле	data.gov.uz, февраль 2023
Дата и охват	февраль 2023, все 6623 набора, CSV без сжатия
Лицензия	не считали
Живая ссылка	каталог скачали целиком; долю ошибок по ссылкам не считали
Общая выгрузка	была, раз файлы удалось забрать
Содержимое	42,6% — одна строка; 96% — меньше 200 строк; 40 МБ CSV
Пять карточек	«Получено яиц…» по Хорезмской области: одна цифра за 2020 год, в мониторинге обновлён 26 декабря 2022
Очередь на месяц	перестать резать показатель по области; мониторинг должен смотреть на строки файла
Очередь на платформу	файл с именем RDF — тот же XML, без словаря

Находка «платформа не умеет то, что помечено обязательным» уходит в карточку решения. Следующий проход проверяет проваленные строки: ушла ли однострочность, открывается ли выгрузка.

Где проход обманывает

- **Свой мониторинг.** Раздел «обновлено» на data.gov.uz ставил дату карточке, в которой одна старая цифра. Свежесть проверяют по файлу.
- **Первая страница и «недавно обновлено».** Там лежат наборы, которые только что трогали.
- **Дата карточки.** Правка описания обновляет `metadata_modified` и не обновляет строки.
- **Число наборов в итоге.** Его поднимают нарезкой. Рядом нужны доля коротких файлов и объём.
- **Сжатое рядом с несжатым.** 29 ГБ выгрузки data.gov.ru и 14 ГБ архива того же снимка — один каталог. 40 МБ CSV и 284 МБ JSON — тоже.
- **Пять карточек одного издателя.** Это диагноз издателя.
- **Зелёный валидатор.** Профиль пройден, файл при этом пустой. Содержимое смотрят отдельно.
- **Доля без слова «выборка».** Тридцать карточек — не перепись каталога из десятков тысяч записей.

Если каталог читают агенты

Если у каталога включён сервер MCP или на него опираются ИИ-агенты, к проходу добавляют ещё один шаг. Его взяли из пилотов Open Knowledge Foundation 2026 года: частые вопросы пользователей становятся тестами.

1. Выпишите десять вопросов, которые чаще всего приходят на почту или в поддержку. Например: «сколько школ в районе», «где расходы бюджета за прошлый год».
2. Для каждого запишите правильный ответ, набор и ссылку на карточку. Число берите из файла, а не из памяти.
3. Задайте те же вопросы агенту, подключённому к серверу каталога.
4. Отметьте три исхода: верное число со ссылкой на набор, верное число без ссылки, неверное или выдуманное.

Неверный ответ чаще всего говорит о пустом описании, а не о модели. В пилоте OKFN модель приписала энергобалансу Уругвая «климатические факторы», которых в наборе нет. Помогли определения полей, единицы, период и известные ограничения в описании. В итог прохода добавляют строку «Ответы агента: верных со ссылкой N из 10». Тест повторяют после каждого обновления сервера и после крупной правки данных. Устройство сервера — в главе об ИИ.

Если портала нет

Страницу, Git-репозиторий и набор в общем хабе проверяют по минимуму публикации без портала: страница открывается без входа, лицензия на месте, разметка schema.org проходит проверку, описание в ней не короче 50 знаков, дата и контакт заполнены. Сверять адрес с реестром каталогов и прогонять валидатор профиля там нечего. Если набор должен попасть в региональный или национальный портал, к минимуму добавляется файл в том профиле, который портал забирает.

Страница итога та же. В полях, для которых нет каталога, пишут «портала нет». Очередь — непройденные пункты минимума.

Источники

- Метрики качества, уровни качества, методики и рейтинги
- Чек-лист требований, чек-лист живого портала
- Антипаттерны, жизненный цикл
- Open Knowledge Foundation, MCP for Open Data Portals: Trust Depends on Understanding the Data, 10 июня 2026 года — частые вопросы партнёров как тесты, «климатические факторы» в ответе об энергобалансе Уругвая
- И. Бегтин, Что не так с порталом открытых данных Узбекистана?, 13 февраля 2023 — 6623 набора, 42,6% из одной строки, 40 МБ CSV, мониторинг
- И. Бегтин, портал открытых данных Кыргызстана, Telegram, 11 февраля 2023 — 646 наборов, около 570 МБ
- И. Бегтин, повторный подсчёт data.egov.uz, Telegram, 10 июля 2025 — 10 412 наборов, 5207 однострочных, JSON без сжатия 426 МБ против 284 МБ в 2023 году

- И. Бегтин, полная выгрузка data.gov.ru, Telegram, 2 февраля 2022 — 26 490 наборов, 29 ГБ без сжатия
- И. Бегтин, Обновлённый портал data.gov.ru, 16 июля 2025 — 5826 наборов, около 100 МБ в ZIP, возраст по первой публикации, архив 2022 года 14 ГБ, ЕМИСС и data.mos.ru

ЧАСТЬ XI.

Тренды

Что меняется в мире каталогов данных.

1. ИИ, пространства данных и регулирование
2. Что уже умеют передовые каталоги

ИИ, пространства данных и регулирование

Каталог, который уже отдаёт описания, файлы, лицензии и полную выгрузку, готов к большинству новинок последних лет. Вход для программы-помощника, поиск фразой, карточка для обучения модели и обмен закрытыми данными между организациями садятся на этот слой. Там, где слоя нет, помощник упирается в капчу, а в пространстве данных не о чем договариваться.

Данные, готовые для ИИ, — набор, который программа может найти, понять и забрать: открытый формат, явная лицензия, происхождение и выгрузка. Отдельной платформы ради этого ярлыка не нужно.

Что добавляют	Кому это помогает	Что должно уже работать
Файл `llms.txt` и сервер MCP	программе, которая ищет наборы в разговоре	выгрузка описаний, API, лицензия на наборе, определения полей и единицы
Поиск фразой	человеку, который не знает точного названия набора	заполненные описания
Карточка Croissant	тому, кто обучает модель, и поисковику наборов	файлы, поля, лицензия, происхождение
Государственная площадка для ИИ	наборам речи, текста и изображений, которые готовят для обучения	это отдельный тип каталога, со своими условиями доступа
Пространство данных	организациям, которые обмениваются данными с разными правами	открытое описание, даже когда файл закрыт
Чат на месте каталога	никому, кому нужно проверить число или загрузить его в свою программу	сам набор, с адресом и лицензией

Программы как пользователи

Программе нужны те же четыре вещи, что и сборщику каталогов: стабильный адрес, полная карточка, явная лицензия, выгрузка без входа. Поисковики данных ходят по каталогам давно. К ним добавились помощники на больших языковых моделях: они ищут набор, читают описание и забирают файл через API.

Файл `llms.txt` — короткая записка о сайте: что здесь опубликовано и куда смотреть. Это предложение формата, а не принятый стандарт. Сервер MCP отдаёт программе готовый список запросов, чтобы она не угадывала адреса. Оба читают то, что каталог уже отдаёт. Пустую карточку такой вход отдаёт быстрее, чем сайт.

Первый заметный вход у национального портала — эксперимент data.gouv.fr. Команда портала объявила его 25 февраля 2026 года. Публичный адрес `https://mcp.data.gouv.fr/mcp` работает только на чтение и без ключа. Сервер ищет наборы, читает карточки, для части таблиц задаёт запрос через API портала и разбирает отдельный файл. В разговор при этом попадает текст и ссылка. Большой файл по-прежнему скачивают обычной выгрузкой. Сам портал предупреждает: рядом ходят серверы, которые называются официальными и ими не являются. Код открыт, лицензия MIT.

Для CKAN готового расширения в поставке нет. Open Knowledge Foundation ведёт раннее исследование: сервер, в котором ответ должен ссылаться на исходный набор. В описании прямо сказано, что проект на ранней стадии и контракт ещё меняется. Независимые серверы

уже садятся на обычный API любого портала CKAN: где API документирован, сервер — тонкая надстройка, а не новая база. Тот же приём встречается и вне открытых данных: у Colectica Repository 8 сервер MCP читает описания микроданных в DDI. Подробнее — в главе о каталогах микроданных.

Для самих таблиц протокол узок. Он рассчитан на короткий ответ, который помещается в разговор. Модель по нему не читает двоичный файл вроде Parquet или GeoPackage. Хороший сервер поэтому возвращает образец, итог запроса и адрес файла. Полную таблицу забирают прежней массовой выгрузкой. Как устроены API и выгрузка — в главе об интерфейсах доступа.

У программ-пользователей есть и обратная сторона. Роботы, которые собирают тексты и файлы для обучения, создают нагрузку, и порталы начинают закрываться. Патентное ведомство США в 2026 году закрыло массовые выгрузки и API портала data.uspto.gov регистрацией. Многие сайты ставят проверку «вы не робот», а государственные сайты ряда стран не отвечают на запросы из-за рубежа, в том числе из облаков, откуда работают сами помощники. Человек из той же страны каталог открывает, программа — нет. От нагрузки защищаются опубликованными лимитами и полной выгрузкой файлом. Подробнее — в главе об антипаттернах.

Сервер MCP: свой, платформенный, чужой

Сервер MCP перестаёт быть самоделкой. Его дают платформы и ставят статистические службы. Вопрос «писать ли свой» сменился вопросом «включать ли готовый и на каких условиях».

Платформы. У Huwise, бывшего Opendatasoft, один сервер на все порталы. В тариф он по умолчанию не входит. Каждый запрос несёт адрес портала и ключ API, без ключа сервер отвечает кодом 401. Агенту доступны поиск по каталогу, в том числе по смыслу фразы, карточка со схемой и образцом записей и запросы к самим записям. Права при этом те же, что у пользователя, вплоть до отдельной колонки. В облаке PortalJS у каждой организации свой адрес сервера, а код сервера открыт. Он ищет наборы, отдаёт карточку со схемой и первые строки таблицы со сводкой колонок. Полный набор в разговор не попадает. Сервер «по умолчанию» — без тарифа, ключа или отдельной установки — у массовых платформ пока редкость.

Статистические службы. По подсчёту Open Data Watch, к августу 2026 года официальные серверы MCP запустили пять национальных статистических служб и агентств данных: США, Великобритании, Франции, Индии и Сингапура. Шестым стал статистический институт Каталонии. Сервер Бюро переписи США ставят у себя: нужны ключ Census API, Docker и локальная база с частью описаний. Индийский сервер портала eSankhyiki в бете и покрывает семь статистических продуктов.

Инструкция для агента. Самый показательный пример — сервер Data360 Всемирного банка. Он отдаёт агенту обязательную инструкцию, справочники кодов и правила поиска. Перед запросом данных инструмент проверяет, есть ли ряд по нужной стране и какие у него годы и

разрезы. Своей базы у сервера нет: каждое значение в ответе берётся из самой платформы. Data Commons раздаёт инструкции для агентов так же, ресурсами своего сервера. Это новый слой описания. Его ведут и проверяют, как карточку.

Что показали пилоты. Open Knowledge Foundation в 2026 году подключала сервер MCP к порталу прозрачности Бразилии и к каталогу открытых данных Уругвая. Вывод в отчёте от 10 июня: подключить просто, понять смысл трудно. В одном ответе модель добавила к энергобалансу Уругвая «климатические факторы», которых в наборе нет. Помогли определения полей, единицы, период, известные ограничения и примеры верных и неверных вопросов. Частые вопросы партнёров превратили в тесты. Это та же работа над описанием, что нужна человеку. Проверяют её на ответах программы.

Чем грозит чужой сервер. Сервер, который ходит к любому порталу по присланному адресу, сам может стать входом в чужую сеть. 21 сентября 2026 года для независимого сервера ondata/ckan-mcp-server до версии 0.4.108 опубликовали уязвимость CVE-2026-61612. Проверка адреса смотрела на строку с именем сервера и не проверяла, куда это имя ведёт. Так можно было обратиться к служебному адресу облака `169.254.169.254`, где лежат его настройки. Это третий обход одной и той же защиты. Свой портал безопаснее открывать сервером, привязанным к одному адресу, только на чтение и с обновлениями от того, кто его ведёт.

Что включать по умолчанию

Если сервер даёт платформа или его ставят сами, у него должны быть шесть свойств.

1. Только чтение.
2. Открытые наборы — без ключа. Ключ нужен только для закрытых.
3. Сервер работает поверх того же публичного API и с теми же лимитами. Своей копии базы у него нет.
4. Ответ ссылается на карточку и файл. Большой файл отдают ссылкой на выгрузку, а не текстом в разговор.
5. Рядом лежит инструкция для агента: определения полей, единицы, период, известные ограничения.
6. Десять частых вопросов с проверенными ответами прогоняют после каждого обновления сервера и данных. Как это вписать в самоаудит — в главе о самоаудите.

Как проверить сервер MCP

Сервер отвечает на обычный запрос POST. Ожидаемые ответы получены 15 октября 2026 года.

```
# Имя и версия сервера, без ключа
curl -s -X POST https://mcp.data.gouv.fr/mcp \
  -H 'content-type: application/json' \
  -H 'accept: application/json, text/event-stream' \
  -d '{"jsonrpc":"2.0","id":1,"method":"initialize","params":{"protocolVersion":"2025-06-
18","capabilities":{},"clientInfo":{"name":"check","version":"1"}}}'

# Список инструментов
curl -s -X POST https://mcp.data.gouv.fr/mcp \
  -H 'content-type: application/json' \
```

```
  -H 'accept: application/json, text/event-stream' \
  -H 'mcp-protocol-version: 2025-06-18' \
  -d '{"jsonrpc":"2.0","id":2,"method":"tools/list"}'

# Инструкции для агента у Data360 лежат ресурсами
curl -s -X POST https://maimcpext.worldbank.org/ext/data360/mcp \
  -H 'content-type: application/json' \
  -H 'accept: application/json, text/event-stream' \
  -H 'mcp-protocol-version: 2025-06-18' \
  -d '{"jsonrpc":"2.0","id":3,"method":"resources/list"}'
```

data.gouv.fr ответил `"name":"data.gouv.fr MCP server","version":"1.28.1"` и десятью инструментами — от `search_datasets` до `query_resource_data` и `get_metrics`. У Data360 те же запросы дают `"Data360 MCP Server"` версии 0.1.0, 16 инструментов и ресурсы `data360://system-prompt`, `data360://codelists`, `data360://search-usage` и другие. Сервер Huwise `https://mcp.huwise.com/mcp` на первый запрос без ключа ответил кодом 401.

Дальше задают серверу три частых вопроса о своих данных. Ответ должен назвать набор и дать ссылку на карточку. Число в ответе сверяют с файлом.

Когда заводить отдельный вход

Контекст. Региональный портал открытых данных на CKAN. Национальный портал забирает описания. Руководство просит «сделать каталог для ИИ»: сервер MCP, поиск фразой и заодно «пространство данных». Выгрузка описаний неполная, у части наборов нет лицензии. Сопровождает один администратор.

Варианты. Починить лицензии и выгрузку. Поставить MCP поверх текущего API. Заменить строку поиска чатом. Вступить в чужое пространство данных или завести своё.

Критерии. Национальный портал продолжает забирать описания. Программа скачивает файл без регистрации. Лицензия видна в карточке и в выгрузке. Новый вход есть кому документировать. Обмен закрытыми данными между организациями либо уже есть в задачах портала, либо его нет.

Решение. Сначала лицензии и выгрузка. MCP ставят после этого, если помощниками уже пользуются и есть кому держать сервер. Если сервер уже даёт платформа, свой не пишут: готовый включают только на чтение и без ключа для открытых наборов. Поиск фразой можно добавить рядом с обычной строкой. Пространство данных не заводят.

Почему. MCP и поиск фразой читают то, что уже отдаёт API. Пустую карточку они отдают быстрее, чем сайт. Чат на месте каталога не оставляет числа, которое можно процитировать и загрузить в свою программу. Пространство данных нужно, когда данными обмениваются многие организации и доступ у них разный. У регионального портала открытых CSV этой задачи нет, а свой коннектор при одном администраторе останется без сопровождения.

Последствия. Помощник, которого подключат позже, увидит те же описания, что и национальный портал. Пока лицензии пустые, набор по-прежнему нельзя спокойно класть в обучающую выборку: неясно, разрешено ли это.

Когда пересмотреть. Появился устойчивый пользователь-программа и человек, который будет обновлять сервер MCP. Платформа начала давать сервер сама: тогда решают не «писать ли свой», а на каких условиях включить готовый. Или несколько организаций просят отдавать им закрытые данные по договору, и заявок больше, чем разбирает почта.

Как записать такую карточку у себя — в главе «Как обосновать решение».

ИИ на стороне каталога

Модель ускоряет черновик описания. Публикует по-прежнему человек.

Самая дорогая часть каталога — тексты карточек. В мае 2025 года реестр каталогов dataportals-registry дополнили с помощью ИИ-агентов: по адресу сайта агент предлагал название, описание, темы и географическую привязку. Ручная работа той же глубины оценивалась в полгода для двух человек. Результат проверяли люди. Так же модель может предложить, что в колонке дата, а что код региона, и набросать словарь полей. Черновик сверяют с файлом: модель уверенно выдумывает и единицы измерения.

Поиск фразой находит набор, а не значение в таблице. На норвежском портале data.norge.no такой поиск стоит рядом с обычной строкой и запросами SPARQL, а не вместо них. О нём писали в апреле 2025 года. Человек спрашивает обычными словами, модель подбирает ссылки на наборы по их описаниям. Сам портал предупреждает, что программный интерфейс этого поиска ещё меняется и покрывает только наборы. Люди при этом часто ищут значение — сколько, где, когда. Такое значение по-прежнему берут из файла или API показателя.

Берлинский Geoexplorer — законченный прототип июля 2024 года поверх метаданных городского геопортала. Он ищет слои фразой и пишет пояснение к набору из выдачи. Код открыт, лицензия MIT, сайт в 2026 году ещё отвечал. Это витрина поиска, а не замена геопортала и не доказательство, что описание слоя можно больше не вести.

Есть и следствие для владельцев данных. Довод «машиночитаемый формат — это дорого» слабеет: если таблица опубликована хоть в каком-то виде, программу разбора можно написать быстро. В ноябре 2025 года стандартный API статистической базы Армении не отвечал, и со страниц собрали больше 900 показателей. К октябрю 2026 года там уже стоял PxWeb выпуска 2025 года, и API заработал. Подробности — в главе о каталогах показателей. Это не повод публиковать сканы. Собранное со страниц ломается при первой переделке вёрстки, а ошибки разбора никто не проверяет.

Что считать готовым для ИИ

Open Data Institute в декабре 2025 года выпустил вторую версию рамки «данные, готовые для ИИ». Она не вводит новый формат каталога. Четыре группы почти целиком уже разобраны в других главах.

Группа в рамке ODI	Что сделать в каталоге	Где разобрано
Сам набор	открытый формат, понятные коды, описанные ограничения	форматы, хорошие практики
Описание	машиночитаемая карточка, лицензия, происхождение	стандарты, лицензии

Группа в рамке ODI	Что сделать в каталоге	Где разобрано
Инфраструктура	портал или страница, API, версии	интерфейсы доступа, идентификаторы
Управление	названный ответственный, контакт, понятный порядок доступа	доступ и рабочие процессы

Тот же институт в ноябре 2024 года проверил, попадают ли британские открытые данные в общие языковые модели. Страницы data.gov.uk в выгрузке Common Crawl за апрель 2024 года нашлись: 13 556 адресов. Это были карточки со ссылкой на скачивание, а не сами таблицы. В проверке на вспоминание модели верно назвали официальную цифру в 5 случаях из 195, и все пять относились к одному набору. Тексты обычных сайтов госорганов модели помнили лучше. Вывод для каталога прямой. Карточка со ссылкой делает набор находимым для человека. В память общей модели число с этой карточки само не попадает. Текущее значение по-прежнему отдают файлом или API, а не ответом модели, обученной в прошлом году.

Государственные каталоги данных для ИИ

Наборы для обучения моделей по-прежнему живут в основном на Hugging Face и Kaggle. Государства рядом строят свои площадки: там другой язык, отраслевые записи и более строгий доступ. Свой портал открытых данных из-за этого не переделывают в ML-каталог. Устройство типа — в главе о ML-каталогах.

- **Индия, AIKosh.** По сообщению IndiaAI от 20 августа 2026 года на платформе было больше 15 тысяч наборов, 300 моделей и 30 инструментов. 26 августа 2026 года на карточках скачивание открывалось после регистрации, у набора стояла оценка качества, а в условиях чаще встречалась лицензия CC BY-NC-SA: коммерческое использование без отдельного разрешения закрыто. Это национальный репозиторий для обучения, со своими правилами на каждом наборе.
- **Корея, AI Hub.** Площадка агентства NIA публикует наборы для обучения: речь, тексты, изображения, в том числе на корейском языке. По правилам площадки описание открыто без входа, а скачивание просит подтвердить личность и назвать цель. Данные предназначены для обучения моделей. Организация за рубежом и вывоз набора за границу требуют отдельной договорённости. Продажа самого набора — тоже отдельный договор, хотя исследования и разработки, в том числе коммерческие, правила в принципе допускают.
- **США, NAIRR.** Пилот национального ресурса для исследований ИИ даёт исследователям доступ к вычислениям, моделям и части наборов. Это не витрина файлов для всех. 1 сентября 2026 года Национальный научный фонд объявил операционный центр, который должен превратить пилот в постоянную службу. Сайт пилота на эту дату ещё принимал заявки. С 2024 года через пилот прошли больше 800 исследовательских проектов.
- **ЕС, AI-on-Demand.** Каталог описаний того, что уже опубликовано на других площадках, включая Zenodo и Hugging Face. 14 октября 2026 года счётчик на главной показывал 118 025 описаний наборов. Это размер указателя, а не архив, который ведёт Еврокомиссия.

Общее требование к такой площадке одно. Карточка называет лицензию на обучение, источник и известные ограничения. Файл может быть закрыт регистрацией. Описание при этом остаётся открытым, иначе набор не найти и нельзя решить, стоит ли подавать заявку.

Дата-продукты

Программе-помощнику нужен набор с известным владельцем, схемой, примерами запросов и стабильным интерфейсом. В разговоре о каталогах это называют дата-продуктом: набор сопровождают, а не выкладывают один раз. Типичный портал открытых данных до такого продукта не дотягивает, поэтому и «семантический слой для ИИ» на нём получается тонким. Чем продукт отличается от набора — в главе о единицах публикации. Практический список — в главе о хороших практиках.

Модель вместо данных

Государство может решить, что открывать таблицы больше не нужно: ответ даст чат, обученный на них. Общая модель этого не умеет. В британской проверке 2024 года она почти не помнила цифры data.gov.uk, хотя страницы портала в обучающую выгрузку попадали. Ответ модели нельзя сверить с файлом, процитировать, сравнить с прошлым годом и загрузить в свою программу. Если открытые данные закрывают и оставляют чат, это закрытие данных.

Чат уместен рядом с каталогом: он подсказывает карточку и ссылается на неё. Число, лицензию и файл чат не заменяет.

Описания для обучения моделей

Набор, на котором сравнивают модели, описывают карточкой Croissant: состав, поля, деление на части, лицензия, известные ограничения. Версия 1.1 вышла 29 января 2026 года. Государственный портал открытых данных из-за этого не обязан становиться Hugging Face. Он обязан так указать лицензию и источник, чтобы набор нельзя было незаметно переложить в обучающую выборку. Как устроен Croissant и кто его уже отдаёт — в главах о стандартах и об ML-каталогах.

Регламент (ЕС) 2024/1689 об искусственном интеллекте эту работу возлагает на поставщика модели, а не на городской портал. С 2 августа 2025 года поставщик модели общего назначения публикует достаточно подробную сводку того, на чём модель учили. У моделей, которые уже были на рынке к этой дате, срок до 2 августа 2027 года. Портал помогает тем, что лицензия и источник стоят в карточке и в выгрузке. Сводку пишет тот, кто учил модель.

Пространства данных

Европейские пространства данных (data spaces) договаривают доступ к данным между организациями, в том числе к данным не для всех. Это ближе к каталогу с уровнями доступа и к маркетплейсу, чем к порталу «скачай CSV».

Общее с порталом открытых данных одно. Описание читается снаружи, иначе договор не о чем заключать. Для этого хватает обычного DCAT. Файл может оставаться закрытым. Условия — «можно для исследований, нельзя передавать дальше» — записывают на языке ODRL. Переговоры о договоре и сама передача идут по Dataspace Protocol: версию 2025-1 Eclipse Foundation утвердил 23 июля 2025 года. Как устроены площадки, коннектор и примеры вроде Catena-X — в главе о других типах каталогов.

Gaia-X в этой конструкции — рамка доверия: кто участник и каким требованиям он соответствует. Это не каталог и не замена описания набора.

Чертёж Data Spaces Support Centre версии 3.0 (февраль 2026 года) собирает, что пространству нужно помимо витрины: схема управления, договоры, общие стандарты. С 2026 года зрелость такой инициативы сверяют с европейской технической спецификацией CEN/CLC/TS 18331:2026, а не с лозунгом на сайте. Европейская комиссия ведёт отраслевые пространства — здоровье, мобильность, промышленность и другие. Городской портал в этот список сам не записывается.

Ситуация	Что делать
Город публикует открытые файлы	аккуратный портал или публикация в региональном
Описание можно показать всем, файл выдают по заявке	каталог с уровнями доступа, как у микроданных
Несколько организаций обмениваются данными, и права у них разные	пространство данных или отраслевая площадка с договорами
Данные продают	маркетплейс; описание всё равно открыто до оплаты
Нужно, чтобы чат отвечал жителям	оставить набор, лицензию и выгрузку; чат ссылается на них

Не каждому городу нужно пространство данных. Большинству по-прежнему нужен портал, в котором описание, лицензия и файл на месте.

Правила

Законы последних лет называют результат публикации. Платформу они называют редко.

Документ	Что меняется для каталога
Директива (ЕС) 2019/1024 об открытых данных	документы госорганов отдают в машиночитаемом виде и по возможности бесплатно; страны переносят это в свои законы
Регламент (ЕС) 2023/138 о наборах высокой ценности, применяется с 9 июня 2024 года	шесть категорий — геоданные, наблюдение Земли и среда, метеорология, статистика, компании, мобильность — бесплатно, через API и массовой выгрузкой
Регламент (ЕС) 2022/868, Акт об управлении данными, применяется с 24 сентября 2023 года	перечень охраняемых данных госсектора: описание открыто, файл выдают по правилам, в том числе в защищённой среде
Регламент (ЕС) 2023/2854, Акт о данных, применяется с 12 сентября 2025 года	доступ к данным подключённых устройств и обмен между компаниями; городскому порталу открытых данных новых CSV обычно не добавляет
Регламент (ЕС) 2024/1689 об искусственном интеллекте	с 2 августа 2025 года поставщик модели общего назначения публикует сводку обучающих данных (у уже выпущенных моделей — до 2 августа 2027 года); портал эту сводку не пишет

Шесть категорий высокой ценности и перечень охраняемых данных разобраны в глоссарии и в главе о доступе. Регламент о Европейском пространстве медицинских данных требует каталоги органов доступа к медицинским данным с 26 марта 2029 года — это тоже каталог с

открытым описанием и закрытым файлом, а не портал открытых CSV.

В США меморандум OMB M-25-05 от 15 января 2025 года — руководство по закону OPEN Government Data Act. Среди целей открытого доступа он называет обучение систем искусственного интеллекта и требует от ведомств инвентаризацию данных и публикацию через федеральный каталог. Обзор Исследовательской службы Конгресса считает меморандум действующим; отчёт той же службы о data.gov от 21 мая 2026 года разбирает его как действующее руководство.

Если закон называет платформу, выбор из этой книги сужается до того, как на этой платформе включить выгрузку, лицензии и постоянные адреса. Если закон называет результат — машиночитаемость, лицензию, бесплатность, API — платформу выбирают как обычно.

Что не изменилось

Постоянный адрес, лицензия на наборе, выгрузка всех описаний и понятный формат файла нужны и человеку, и поисковику, и агенту. Каталог, который это уже делает, переживает смену моды без переписывания. Каталог-витрина не переживает: его не видит ни федерация, ни программа.

Перед новым входом стоит проверить шесть вещей.

1. Программа забирает все описания без входа и без капчи.
2. Лицензия стоит на наборе и попадает в выгрузку.
3. Файл скачивается по постоянной ссылке, а не только открывается в чате.
4. Поиск фразой, если его включили, возвращает карточки. Обычная строка остаётся.
5. Ответ чата, если он есть, ссылается на набор. Набор при этом по-прежнему можно скачать.
6. Сервер MCP, если он включён, работает только на чтение, а на частые вопросы отвечает проверенными числами со ссылкой на набор.

Что из новых возможностей каталогов стоит заказывать, а что пока только чужой сервис, — в главе о передовых каталогах.

Источники

- Croissant 1.1, 29 января 2026 года
- data.gouv.fr, объявление об эксперименте MCP, 25 февраля 2026 года; руководство; код datagouv/datagouv-mcp, лицензия MIT
- Open Knowledge Foundation, mcp-server — раннее исследование, контракт ответов ещё меняется; независимый сервер для API CKAN — ondata/ckan-mcp-server. Проверено 14 октября 2026 года
- Open Knowledge Foundation, MCP for Open Data Portals: Trust Depends on Understanding the Data, 10 июня 2026 года — пилоты в Бразилии и Уругвае, частые вопросы как тесты
- NVD, CVE-2026-61612, 21 сентября 2026 года — ondata/ckan-mcp-server до 0.4.108, проверка адреса без разрешения имени, третий обход защиты

- Huwise, сервер MCP — не входит в тариф автоматически, ключ API и адрес портала в каждом запросе; PortalJS, сервер MCP, 25 ноября 2025 года, код datopian/portaljs-mcp-server
- Open Data Watch, Making Official Statistics Work in an AI-Driven World, сентябрь 2026 года — пять официальных серверов MCP национальных служб к августу 2026 года
- Всемирный банк, Data360 MCP — инструкция для агента, справочники кодов, проверка покрытия до запроса данных; Data Commons, работа с ИИ-агентом
- U.S. Census Bureau, Data API MCP, лицензия CC0; PIB, сервер MCP статистической службы Индии, 6 февраля 2026 года
- Open Data Institute, The UK government as a data provider for AI, ноябрь 2024 года — 5 верных ответов из 195 и 13 556 страниц data.gov.uk в Common Crawl за апрель 2024 года; A framework for AI-ready data, версия 2, декабрь 2025 года
- data.norge.no: поиск с ИИ и описание API; новость data.europa.eu, 25 апреля 2025 года
- Geoexplorer Berlin, карточка проекта (июль 2024 года, статус «завершён»), код под лицензией MIT
- IndiaAI / PIB, семинар по AIKosh, 20 августа 2026 года — больше 15 тысяч наборов, 300 моделей, 30 инструментов; И. Бегтин, разбор карточек AIKosh, Telegram, 26.08.2026
- Правила использования AI Hub — подтверждение личности, цель скачивания, отдельное соглашение для использования из-за рубежа
- NAIRR Pilot; NSF, операционный центр, 1 сентября 2026 года — больше 800 проектов с 2024 года
- AI-on-Demand — 14 октября 2026 года счётчик показывал 118 025 описаний наборов; в выдаче каталога рядом стоят Zenodo и Hugging Face
- Data Spaces Support Centre, Blueprint 3.0, февраль 2026 года; сводка; CEN/CLC/TS 18331:2026 — спецификация оценки зрелости, одобрена в апреле 2026 года; Европейская комиссия, общие европейские пространства данных
- Регламент (ЕС) 2024/1689 об искусственном интеллекте — сводка обучающих данных у поставщика модели общего назначения с 2 августа 2025 года
- OMB, меморандум M-25-05, 15 января 2025 года; Congressional Research Service, IN12502 и R48954, 21 мая 2026 года
- И. Бегтин, ИИ и открытые данные. Мнения и сомнения, март 2026 — MCP у порталов и риск замены данных моделью
- И. Бегтин, Telegram: ИИ-агенты в реестре каталогов, 14.05.2025, и итог обогащения, 22.05.2025; удешевление сбора данных, 05.01.2026; геоблокировки и ИИ-агенты, 14.07.2025; data.uspto.gov, 23.06.2026; MCP для данных и двоичные форматы, 03.07.2025
- Директива (ЕС) 2019/1024, регламент (ЕС) 2023/138, регламент (ЕС) 2022/868, регламент (ЕС) 2023/2854
- Находимость, маркетплейсы и каталоги API

Что уже умеют передовые каталоги

Передовые каталоги отличаются не платформой, а тем, что сокращают путь от карточки до ответа. Одни сами превращают файл в API. Другие выдают DOI на выборку, хранят версии таблицы как код, проверяют набор до публикации или считают на своей стороне. Почти всё это пока не входит в ядро CKAN, ArcGIS Hub, Huwise, Socrata, Dataverse, DSpace и GeoNetwork. Часть можно поставить рядом со своим порталом: есть открытый код. Часть пока доступна только как чужой сервис. От этого и зависит, что из главы писать в техническое задание.

Ниже девять таких возможностей. У каждой — один-два каталога, на которых её видно лучше всего, и ответ на три вопроса: чем повторить, когда не нужно и как проверить запросом.

Каталоги отобраны по четырём признакам:

- возможности нет в ядре массовых платформ: то, что включается настройкой, — уже не новшество;
- она работает публично, и её можно проверить запросом без входа;
- идею можно повторить: есть открытый код, стандарт или хотя бы описанное соглашение;
- понятен статус. Бета, единственный оператор или обещание хранить пять лет — часть описания, а не сноска.

Возможность	Где видно лучше всего	Чем повторить у себя	Когда не нужно
API к каждому табличному файлу	data.gouv.fr, Hugging Face, Datasette	открытый сервис api-tabular, Datasette; в CKAN — DataStore с загрузчиком	файлов мало, они маленькие и их скачивают целиком
Идентификаторы объектов внутри данных	Overture Maps, OpenSanctions	правила идентификаторов в реестре, таблицы соответствия	набор ни с чем не связывают
DOI на выборку	GBIF; проще — Our World in Data, CELLxGENE Census	DOI от DataCite на сохранённый запрос; соглашение об адресах	на данные не ссылаются
Версии таблиц как в Git	DoltHub, OpenNeuro	Dolt и DoltLab, DataLad	таблицу правит один человек раз в год
Проверка по стандарту на входе	OpenNeuro, DANDI	схема и валидатор при загрузке	у области нет общего стандарта
Обработка рядом с данными	Copernicus Data Space Ecosystem, Climate Data Store, ERDDAP	ERDDAP, сервер openEO	данные помещаются в обычный файл
Сведение многих источников	Data Commons, HDX HAPI, OpenSanctions	открытый код Data Commons, yente	каталог одной организации
Доступ по ступеням	microdata.no, Redivis	своя разработка или чужой сервис	данные открыты целиком
Готовность к ИИ-агентам	data.gouv.fr, Data360 Всемирного банка	сервер MCP, который даёт платформа	нет лицензий и полной выгрузки

С чего начинается ряд: Source Cooperative

В Source Cooperative ново не хранилище, а то, что каталог почти не стоит между читателем и данными. Продукт — папка в объектном хранилище с описанием README. Файлы в облачных форматах читают по частям прямо по сети. В реестре dataportals-registry у площадки записан интерфейс STAC — протокол описания пространственных файлов. Сильные и слабые стороны — бета, публикация по заявке, нет DOI и версий — разобраны в главе об облачных

хранилищах и хабах. Следующие два раздела развивают ту же идею. У Overture Maps постоянные идентификаторы есть у объектов внутри данных. У Hugging Face по файлу сразу можно выполнить запрос.

API к каждому табличному файлу

Обычно между «скачать файл» и «получить API» стоит работа издателя: загрузить таблицу в базу, описать колонки, поддерживать загрузчик. Передовые каталоги делают это сами.

data.gouv.fr. Табличный API французского портала (пока в бете) сам подхватывает каждый табличный файл в течение нескольких минут после публикации. Предел по размеру: 100 МБ для CSV, 50 МБ для XLS и Parquet, 12,5 МБ для XLSX. У каждого файла три адреса. Профиль показывает типы колонок и распознанные форматы — например, код коммуны INSEE или широту, — частые значения и число пропусков. Данные отдаются постранично, с фильтрами и сортировкой. Описание запросов в формате Swagger строится для каждого файла отдельно. Предел — 100 запросов в секунду. На странице сервиса к нему привязаны 31 768 наборов. Код сервиса открыт под лицензией MIT, его можно поставить рядом со своим порталом.

Hugging Face. Платформа сама переводит наборы поддерживаемых форматов в Parquet и строит по ним просмотр, поиск, фильтры и статистику. Консоль SQL работает целиком в браузере на DuckDB. Запросом можно поделиться ссылкой, а результат выгрузить в Parquet или CSV. Это сдвигает границу, которую книга проводит в разделе о предпросмотре: предпросмотр нельзя процитировать, а ссылка на запрос уже воспроизводима. Постоянной ссылкой она всё же не становится. Набор могут обновить, и тот же запрос вернёт другие строки.

Datasette — тот же принцип в малом масштабе. У любой страницы есть версия в JSON, фильтры пишутся в адресе, есть полнотекстовый поиск и произвольный SQL. В реестре 45 записей на Datasette, 44 из них действующие. Как опубликовать на нём небольшой набор, — в главе о статических сайтах.

Чем повторить. В CKAN тот же результат дают DataStore и расширение XLoader, но предел строк и загрузчик остаются на вашей стороне. Подробнее — в главах об интерфейсах доступа и о хранилище.

Когда не нужно. Если файлов десяток и их скачивают целиком, API ничего не добавит. Ограничения надо называть честно. Большие файлы остаются просто файлами. Бета может поменять адреса. API к чужому CSV наследует все ошибки этого CSV: неверную кодировку, лишние строки над заголовком, текст в числовой колонке.

Идентификаторы объектов внутри данных

Книга подробно разбирает DOI набора. Но идентификатор может быть и у отдельной строки. Это меняет то, как данные связывают между собой.

Overture Maps. У каждого здания, дороги, места и административной единицы в ежемесячном выпуске Overture есть постоянный идентификатор в системе GERS. Рядом ведутся реестр идентификаторов с датами первого и последнего появления, журнал изменений по каждому объекту и таблицы соответствия с идентификаторами исходных

источников. Зачем это нужно, Overture объясняет так. Поставщик данных о пробках может публиковать только пробки, без самих дорог, а поставщик показателей — только показатели, без границ. Соединяют их по идентификатору, без сопоставления геометрий.

Здесь же честный урок. В выпуске 2025-06-25.0 Overture заменил старые идентификаторы, в которых были зашиты ячейка сетки H3 и тема, на случайные UUID. Старые идентификаторы больше не находятся. Идентификатор, в который зашит смысл, обходится дороже, чем кажется: смысл меняется, а ссылки на старый номер остаются.

OpenSanctions показывает ту же идею для людей и компаний. Сущность собирают из многих списков, после устранения дублей у неё один канонический идентификатор. Если она совпала с записью Викиданных, идентификатором становится номер Викиданных.

Чем повторить. Правила реестра: идентификатор по правилам, а не номер строки; записи не удаляются, а получают статус; таблица соответствия со старыми номерами. Они уже перечислены в главе об идентификаторах.

Когда не нужно. Набор, который никто не связывает с другими, — годовой отчёт, разовое обследование, — обходится идентификатором набора.

DOI на выборку

Обычно DOI выдают набору. Передовые каталоги выдают его запросу.

GBIF. У каждой выгрузки наблюдений есть DOI. Страница выгрузки хранит список наборов-источников и все условия отбора. 15 октября 2026 года набор наблюдений iNaturalist входил в 2 623 241 выгрузку, и у каждой был свой DOI. Если исследователь потом отфильтровал данные, собрал их из нескольких выгрузок или посчитал в облаке, он регистрирует производный набор со списком исходных и тоже получает DOI. Ежемесячные снимки GBIF в облаках Microsoft, Amazon и Google имеют DOI. GBIF отслеживает, где эти DOI процитированы. Ссылки на выгрузку разобраны в главе о цитировании. Новое здесь не то, что DOI есть, а то, что объектом публикации становится запрос.

Рядом стоят два решения попроще.

Our World in Data позволяет дописать к адресу любого графика `.csv`, `.metadata.json` или `.zip` и получить данные, описание и README. В описании есть готовая строка источников для ссылки. Это соглашение об адресах, его можно перенять без какого-либо ПО.

CELLxGENE Census публикует данные одноклеточного секвенирования как один объект, из которого берут срезы по клеткам и генам. Выпуски с долгой поддержкой обещаны раз в полгода, каждый хранится не меньше пяти лет. Для любого среза есть генератор ссылки. Обещание хранить выпуск с названным сроком — хороший образец для главы об архивации. Статус при этом сверяют: 15 октября 2026 года файл выпусков называл стабильным выпуск 2025-11-08, а последним — 2025-11-17.

Чем повторить. DOI на сохранённую выборку выдают через DataCite так же, как на набор. Нужны сохранённый текст запроса, дата и список источников. Общая схема описана в рекомендациях Research Data Alliance о ссылках на меняющиеся данные.

Когда не нужно. Если на данные не ссылаются в работах, хватает даты обращения и архива прошлых выгрузок.

Версии таблиц как в Git

Dolt — база данных с языком SQL, совместимая с MySQL. Её можно клонировать, ветвить, сливать и отправлять так же, как репозиторий Git. **DoltHub** бесплатно хранит публичные базы. Изменение данных проходит через запрос на слияние с проверкой: проверяющий видит разницу по строкам и ячейкам, а не по файлу. DoltLab — та же площадка для установки у себя. Это вариант между Git-репозиторием и порталом: история правок как в Git, но для таблицы, которая уже не помещается в CSV на 2 МБ. Подробнее — в главе о Git-репозиториях.

OpenNeuro — та же идея в научном архиве. Каждый снимок набора — версия со своим хэшем Git и DOI. Набор можно забрать из S3, через DataLad или прямо из Git. DataLad — программа, которая хранит большие наборы как репозитории Git: история и описание в Git, сами файлы — отдельно. В реестре на DataLad 10 установок.

Когда не нужно и чем рискует. Таблицу, которую раз в год правит один человек, проще держать в обычном Git. Площадка с одним оператором может закрыться: в 2022 году так закрылся каталог QRI с той же идеей, история — в главе об облачных хранилищах и хабах. Поэтому в карточке решения нужен ответ, как забрать базу и историю правок.

Проверка по стандарту на входе

Часть книги о качестве измеряет его после публикации. Архивы нейроданных проверяют его до публикации.

OpenNeuro принимает только наборы по стандарту BIDS — общему для нейровизуализации порядку папок, имён файлов и описаний. При загрузке набор проверяет валидатор. С предупреждениями загрузить можно, с ошибками нельзя. Условия жёсткие: снимки обезличены, лицензия CC0 сразу или через 36 месяцев после первого снимка.

DANDI — архив нейрофизиологии на стандарте NWB. Кнопка публикации становится активной только после того, как исправлены все ошибки проверки описания набора и файлов. Опубликованная версия не меняется и цитируется по номеру. Ограничение описано честно: наборы с файлами Zarr пока опубликовать нельзя. В реестре три установки DANDI.

Здесь качество — условие входа, а не показатель на дашборде. Это работает там, где у области есть общий стандарт. Для городского портала аналог — проверка таблицы по схеме из каталога schema.data.gouv.fr при загрузке. Как это вписать в уровни качества — в главе об уровнях качества.

Когда не нужно. Если общего стандарта нет, проверка на входе превращается в список правил одного портала. Издатели обходят его пустыми полями.

Обработка рядом с данными

Здесь каталог отдаёт не файл, а нужный кусок данных или результат расчёта.

Copernicus Data Space Ecosystem даёт к одним и тем же спутниковым данным S3, STAC, openEO и Sentinel Hub. openEO — открытый протокол, по которому пользователь отправляет на сервер цепочку операций над снимками: вырезать участок, посчитать индекс растительности, усреднить по месяцам. Считает сервер, пользователь получает результат. Есть федерация с партнёрскими платформами и отдельная точка доступа без неё. Пределы бесплатного уровня названы прямо: два синхронных запроса и два пакетных задания одновременно, проверено на участках до 100×100 км при разрешении 10 м. В реестре 12 записей с openEO в адресе, три из них — у Copernicus Data Space Ecosystem.

Climate Data Store решает проще и для человека. Пользователь собирает выборку в форме на странице набора, а кнопка «Show API request code» выдаёт готовый код запроса на Python. Условия использования при этом принимают вручную на странице набора, один раз.

ERDDAP напоминает, что новшество не обязано быть новым. Весь запрос, вместе с выборкой по пространству и времени в обычных единицах, умещается в адрес. Формат ответа выбирают на лету: CSV, JSON, NetCDF, MATLAB, KML, PNG. Есть подписка на изменения набора. В реестре 175 записей на ERDDAP, 174 из них действующие. Это больше, чем у любого другого решения из этой главы.

Когда не нужно и чем рискует. Если данные помещаются в обычный файл, вычисления на сервере не нужны. Вычисления дороже всего, и их закрывают первыми: в 2024 году Microsoft закрыла Planetary Computer Hub, а каталог остался. Поэтому каталог и вычисления описывают в требованиях отдельными пунктами. Потеря второго не должна означать потерю первого.

Сведение многих источников

Обычный каталог знает файлы. Эти каталоги знают переменные, места и сущности и отвечают на вопрос сразу по всем источникам.

Data Commons собирает статистику многих организаций в граф знаний — одну базу, где у каждой переменной и каждого места свой идентификатор, а связи между ними записаны явно. Данные отдаются через REST, SDMX 3.0, BigQuery, встраиваемые компоненты и сервер MCP. «Свой Data Commons» можно развернуть на открытом коде.

HDX HAPI особенно полезен читателю с CKAN. Гуманитарный портал HDX работает на CKAN, а API CKAN отдаёт только описания. HAPI добавляет сверху приведённые к одному виду значения показателей из многих источников: население, продовольственная безопасность, финансирование. Обновляются они ежедневно, сами числа не меняются. Это слой поверх существующего портала без смены платформы.

OpenSanctions хранит каждое значение как отдельное утверждение: из какого источника, какое свойство, когда оно впервые и в последний раз встречалось, с каким каноническим идентификатором сущность слилась после устранения дублей. Таблица утверждений выгружается каждую ночь. Это происхождение на уровне значения, а не набора. Сервис поиска и сопоставления yente открыт под MIT. Сами данные бесплатны только для некоммерческого использования.

Рядом стоят уже описанные в книге DBnomics — один API ко многим поставщикам статистики с сохранёнными исходными кодами — и IPUMS — конструктор выгрузок и единый API для всех коллекций микроданных. Подробнее — в главах о каталогах показателей и о каталогах микроданных. У **Aleph** пользователь загружает свой список людей или компаний, и система сверяет его с сотнями наборов, к которым у него есть доступ. Признак тревожный: последний коммит в основном репозитории Aleph был 20 февраля 2026 года.

Когда не нужно. Каталогу одной организации сводить нечего. Сведение нужно агрегатору и тому, кто отвечает на вопросы сразу по многим источникам.

Доступ по ступеням

microdata.no — самый сильный пример. Сотрудники и студенты одобренных вузов, научных организаций и министерств работают с регистровыми данными Норвегии без заявки и без заранее описанного проекта. Записей они не видят. Данные не покидают защищённое хранилище, а результаты обезличиваются на лету.

Redivis делит доступ к набору на пять ступеней: набор невидим, видно описание, видны переменные и сводная статистика, доступна выборка-образец, доступны все данные. Каждую ступень можно открыть всем, отдельным людям или после выполнения условий. В реестре 13 записей на Redivis и 27 каталогов с интерфейсом GA4GH DRS — стандартом выдачи файлов биомедицинских данных по правам доступа.

Ступенчатый доступ — функция каталога. Описание и статистика открыты всем, данные выдаются по ступеням. Читателю с чувствительными данными это даёт вариант между «всё открыть» и «ничего не показывать». Как устроены уровни доступа в массовых платформах — в главах о доступе и ролях и о каталогах микроданных.

Когда не нужно. Если данные открыты целиком, ступени только прячут описание за лишним шагом.

Готовность к ИИ-агентам

ИИ-агент — программа на большой языковой модели, которая сама ищет набор, читает описание и забирает данные. Сервер MCP даёт такому агенту готовый список запросов к каталогу. Подробно об этом — в главе об ИИ и пространствах данных. Здесь — что в этой области уже считается передовым.

Сервер теперь дают сами платформы, но пока почти никто не включает его по умолчанию. У Huwise сервер есть, но в тариф он не входит, и каждый запрос требует ключа API. У каждой организации в облаке PortalJS свой адрес сервера, а его код открыт. data.gouv.fr держит свой сервер только на чтение и без ключа. Лучший пример устройства — сервер Data360 Всемирного банка. Кроме инструментов он отдаёт агенту обязательную инструкцию, справочники кодов и проверку того, есть ли ряд по нужной стране, до запроса данных. Data Commons раздаёт инструкции для агентов так же, ресурсами своего сервера. Это новый слой описания: инструкция к данным для программы, которую тоже надо вести.

Когда не нужно. Пока у наборов нет лицензий и полной выгрузки, сервер быстрее отдаёт пустые карточки. Своими силами его ставят после этого. Если сервер даёт платформа, его включают только на чтение, без ключа для открытых наборов и поверх того же публичного API.

Что не выделять как новшество

Чат на месте каталога и поиск фразой уже разобраны в главе об ИИ, вместе с риском «модель вместо данных». Машинный перевод карточек на 24 языка ЕС у data.europa.eu — в главе о других возможностях. Открытый код витрины статистики Малайзии OpenDOSM и data.gov.my — в главе о каталогах показателей. Небольшие статические порталы «данные как код» на Datadex в реестре представлены тремя записями, и ни у одной API не отмечен. Это пока эксперимент.

Как проверить

Команды не требуют входа. Ожидаемые ответы получены 15 октября 2026 года.

```
# data.gouv.fr: профиль, строки и описание Swagger у одного CSV.
# RID — идентификатор файла из карточки набора
RID=571a3021-0e10-471c-98b8-72bcab5daca8
curl -s "https://tabular-api.data.gouv.fr/api/resources/$RID/profile/"
curl -s "https://tabular-api.data.gouv.fr/api/resources/$RID/data/?
page_size=2&commune__exact=Cessy"
curl -s -o /dev/null -w "%{http_code}\n" "https://tabular-
api.data.gouv.fr/api/resources/$RID/swagger/"

# Hugging Face: готовые файлы Parquet и что включено у набора
curl -s "https://datasets-server.huggingface.co/parquet?dataset=nyu-mll/glue"
curl -s "https://datasets-server.huggingface.co/is-valid?dataset=nyu-mll/glue"

# Overture Maps: каталог выпусков в STAC
curl -s "https://stac.overturemaps.org/catalog.json"

# GBIF: в скольких выгрузках с DOI есть набор iNaturalist
curl -s "https://api.gbif.org/v1/occurrence/download/dataset/50c9509d-22c7-4a22-a47d-
8c48425ef4a7?limit=1"

# Our World in Data: данные и описание графика по соглашению об адресах
curl -s -o /dev/null -w "%{http_code} %{content_type}\n"
"https://ourworldindata.org/grapher/life-expectancy.csv"
curl -s "https://ourworldindata.org/grapher/life-expectancy.metadata.json"

# CELLxGENE Census: выпуски и пометка долгой поддержки
curl -s "https://census.cellxgene.cziscience.com/cellxgene-census/v1/release.json"

# DoltHub: SQL к истории правок базы
curl -s "https://www.dolthub.com/api/v1alpha1/dolthub/ip-to-country/master?
q=SELECT%20commit_hash%2C%20committer%2C%20date%20FROM%20dolt_log%20LIMIT%202"

# OpenNeuro: последний снимок набора
curl -s https://openneuro.org/crn/graphql -H 'content-type: application/json' \
  -d '{"query":"{ dataset(id:\"ds000001\"){ latestSnapshot { tag created } } }"}'

# DANDI: версии набора и итог проверки черновика
curl -s "https://api.dandiarchive.org/api/dandisets/000003/versions/?page_size=3"
```

```
# Copernicus Data Space Ecosystem: сервер openEO
curl -s "https://openeo.dataspace.copernicus.eu/openeo/1.2/"

# HDX HAPI: ключ приложения выдаётся сразу, без регистрации
curl -s "https://hapi.humdata.org/api/v2/encode_app_identifier?
application=book&email=book%40example.org"

# OpenSanctions: когда обновлён основной набор
curl -s "https://data.opensanctions.org/datasets/latest/index.json"
```

Что было в ответах:

- data.gouv.fr: профиль распознал в колонках `lat` и `lon` широту и долготу (`latitude_wgs`, `longitude_wgs`), фильтр вернул строку коммуны Cessy, описание Swagger ответило кодом 200. Файл взят из набора о доступности слухопротезистов по коммунам. Если набор снимут, подставьте идентификатор любого CSV до 100 МБ;
- Hugging Face: адреса файлов Parquet в ветке `refs/convert/parquet`; у набора включены просмотр, поиск, фильтры и статистика;
- Overture: каталог выпусков на STAC 1.1.0, среди выпусков 2026-08-19.0 и 2026-09-23.0;
- GBIF: `count` — 2 623 241, у первой выгрузки в ответе свой DOI, лицензия и условия отбора;
- Our World in Data: файл CSV с кодом 200, в описании строка источников «Riley (2005); Zijdeman et al. (2015); HMD (2025); UN WPP (2024)»;
- CELLxGENE Census: `"stable": "2025-11-08"`, `"latest": "2025-11-17"`, у выпусков с долгой поддержкой пометки `"lts": true` и `"do_not_delete": true`;
- DoltHub: `"query_execution_status":"Success"`. Ветка у этой базы называется `master`, по адресу с `main` ответ — «branch not found»;
- OpenNeuro: у `ds000001` снимок `1.0.0` от 14 мая 2020 года;
- DANDI: `count` — 5, первая версия в списке — черновик `draft` со статусом `Valid`;
- openEO: `api_version` — 1.2.0;
- HDX HAPI: строка `encoded_app_identifier`, её передают в параметре `app_identifier` следующих запросов;
- OpenSanctions: у набора `default` дата обновления `2026-10-15T12:54:35`.

Как проверить сервер MCP, — в главе об ИИ. Цифры реестра получены запросами из приложения «Как работать с реестром».

Что из этого следует

Девять возможностей — не список обязательного. Это список того, что можно попросить, если есть задача. Чтобы он работал как инструмент выбора, у каждой возможности в техническом задании должна быть пометка «обязательно», «желательно» или «не нужно» и фраза, откуда потребность. Готовые формулировки — в разделе «Новые возможности» чек-листа требований.

Три вещи из главы годятся почти любому каталогу:

- API к табличным файлам, если файлы небольшие и их правда фильтруют;
- проверка по схеме при загрузке, если схема уже есть;

- сервер MCP только на чтение, если его даёт платформа.

Остальное нужно под задачу: DOI на выборку — тем, на чьи данные ссылаются; ступени доступа — владельцу чувствительных данных; обработка на сервере — тем, у кого данные не помещаются в файл.

Источники

- dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года — ERDDAP 175 записей (174 действующие), Datasette 45 (44), Redivis 13, DataLad 10, DANDI 3, Datadex 3 (API не отмечен ни у одной), 27 каталогов с GA4GH DRS, 12 записей с openEO в адресе
- data.gouv.fr, API tabulaire (beta) — пределы размера, профиль, 100 запросов в секунду, 31 768 наборов; код datagouv/api-tabular, лицензия MIT
- Hugging Face, Dataset viewer и SQL Console
- Datasette, JSON API
- Overture Maps, GERS — реестр идентификаторов, журнал изменений, таблицы соответствия, смена идентификаторов в выпуске 2025-06-25.0
- GBIF, рекомендации по цитированию — DOI выгрузок, производные наборы, облачные снимки
- Our World in Data, Chart API
- CZ CELLxGENE, Census — выпуски с долгой поддержкой, генератор ссылки
- Research Data Alliance, Precise, Actionable Reference to Dynamic Data, 2025 — ссылки на меняющиеся данные
- DoltHub, What is Dolt?, SQL API
- OpenNeuro, руководство пользователя — BIDS, снимки, CC0 через 36 месяцев; BIDS
- DANDI, публикация наборов — проверка перед публикацией, неизменная версия, ограничение для Zarr
- Copernicus Data Space Ecosystem, openEO — федерация и пределы бесплатного уровня
- Climate Data Store, настройка API
- ERDDAP, описание
- Data Commons, работа с ИИ-агентом
- HDX, Humanitarian API
- OpenSanctions, модель утверждений; yente, лицензия MIT
- DBnomics; IPUMS, API Program
- Aleph, перекрёстная сверка; репозиторий — последний коммит 20 февраля 2026 года
- microdata.no; Redivis, уровни доступа
- Huwise, сервер MCP; PortalJS, сервер MCP; Всемирный банк, Data360 MCP
- Облачные хранилища и хабы, ИИ и пространства данных, чек-лист требований

Приложения

Глоссарий, работа с реестром, сравнительные таблицы, чек-листы и литература.

1. Глоссарий
2. Как работать с реестром
3. Сравнительные таблицы платформ
4. Чек-листы
5. Литература и ссылки
6. Стабильные адреса для ссылок

Глоссарий

Автоматический API к таблице — каталог сам разбирает опубликованный табличный файл и отдаёт к нему API: строки, фильтры, профиль колонок. Издатель ничего не загружает. Пример — табличный API data.gouv.fr. Ошибки исходного файла такой API наследует. Подробнее — в главе о передовых каталогах.

Агрегатор данных — сервис, который собирает метаданные из многих каталогов и даёт по ним единый поиск. У агрегатора у каждого собранного набора своя страница, которую видят обычные поисковики; поисковик данных такой страницы не заводит и сразу отправляет к источнику.

Административные данные — данные, которые возникают как побочный продукт работы ведомства: реестры лицензий, контракты, проверки, платежи. Охват у них обычно полный, но поля придуманы для учёта, а не для анализа.

Аутентификация и авторизация — аутентификация проверяет, кто пользователь: пароль, второй фактор, учётная запись организации или государственная система входа. Авторизация проверяет, что ему можно: роль, права на набор, одобренная заявка на доступ. Единый вход переносит в чужую систему только первое, права выдаёт сам каталог.

Балл доверия — необязательная оценка каталога в реестре dataportals-registry, от 0 до 100. Складывается из типа владельца, типа каталога, лицензии, карточки в re3data и того, отвечает ли сайт. Это фильтр для тысяч чужих каталогов. Сертификатом он не является и файлы не открывает.

«Безголовая» схема (headless) — платформа каталога работает только как ядро: хранит описания, отдаёт API, собирает чужие каталоги, а сайт для людей делают отдельно. Так устроены, например, порталы Торонто и Канады поверх CKAN.

Бенчмарк — набор задач и данных, на которых разные модели машинного обучения сравнивают по одним правилам.

Битая (нерабочая) ссылка — ссылка на файл или страницу, которая больше не ведёт к данным: сервер отвечает «не найдено», домен не существует, сертификат не принимается. Опаснее «мягкая» поломка: ссылка открывается, но ведёт на главную страницу сайта или на страницу вместо файла, и программа, которая смотрит только на код ответа, считает её рабочей.

Веб-архив — собрание снимков сайтов, сделанных краулером в разные даты. Крупнейший — Интернет-архив (web.archive.org), свои веб-архивы ведут национальные библиотеки и архивы. Страницы сохраняет хорошо, данные из программного интерфейса и большие файлы — плохо.

Версия и ревизия — версия набора — зафиксированное состояние данных, на которое можно сослаться; в научных репозиториях у каждой версии свой идентификатор. Ревизия — мелкая правка описания внутри одной версии, новой версии она не создаёт.

Владелец данных — человек или подразделение, которые отвечают за строки набора: верны ли они, когда их обновлять и кто ответит на ошибку. Издатель в карточке — организация. Владелец может быть конкретным человеком внутри неё.

Внешний инструмент — отдельная программа, которую каталог открывает для просмотра или обработки файла. Код программы в каталог не входит: каталог хранит адрес и передаёт идентификатор файла. У Dataverse так устроены просмотрщики.

Возможность каталога — то, что каталог умеет делать для человека, программы или владельца: отдать описания в DCAT, выдать DOI, закрыть файл при открытом описании. Одна и та же возможность бывает в ядре платформы, в расширении или отсутствует.

Воспроизводимость — возможность заново получить заявленный результат по выложенным данным и записанным шагам. Значок «открытые данные» и DOI говорят, что файл выложен и на него можно сослаться. Совпадает ли расчёт с заявленным результатом, проверяют отдельно.

Временной ряд — значения одного показателя при одних и тех же разрезах, расположенные по времени: курс доллара к евро по дням, население региона по годам. В хорошем каталоге показателей у ряда есть постоянный адрес, собранный из кодов справочников.

Выпуск с долгой поддержкой (LTS) — выпуск данных, который обещано хранить без изменений названный срок. У CELLxGENE Census это не меньше пяти лет. Срок записан заранее, а не «пока работает сервер».

Гармонизация микроданных — приведение переменных разных переписей и обследований к одним кодам и определениям, чтобы страны и годы можно было сравнивать. Так работает IPUMS.

Геоблокировка — ограничение доступа к сайту для посетителей из других стран по их IP-адресу. Вместе с иностранными пользователями сайт перестают видеть веб-архивы, поисковики и агрегаторы данных.

Геопортал — каталог пространственных данных и сервисов, часто совмещённый с веб-картой.

Гибрид (гибридный каталог) — сайт, на котором рядом живут продукты разных типов, например наборы и слои карт, или организация, которая ведёт каталоги нескольких типов. Не отдельный тип: каждую часть гибрида классифицируют и оценивают по её собственному типу. Самый развитый вариант — хаб данных.

Граф знаний — база, где у каждой переменной, места или сущности свой идентификатор, а связи между ними записаны явно. Позволяет отвечать на вопрос сразу по многим источникам. Так устроен Data Commons.

Данные как продукт (дата-продукт) — подход, при котором у набора есть владелец, известные пользователи, график обновления, документация и контроль качества, как у любого продукта; доступ обычно устроен несколькими способами сразу — API, выгрузка, срезы. Противоположность — файл, выложенный один раз «для галочки».

Данные, готовые для ИИ — набор, который программа может найти, понять и забрать: открытый формат, явная лицензия, происхождение и выгрузка. Карточка Croissant и сервер MCP это дополняют, когда формат и лицензия уже есть. Рамка Open Data Institute (версия 2, декабрь 2025 года) раскладывает то же на свойства набора, описание, инфраструктуру и управление.

Дистрибутив (способ получения) — конкретная форма набора данных: файл, ссылка, API, сервис. Термин из DCAT. В CKAN называется ресурсом, в Dataverse и Zenodo — файлом, в STAC — ассетом. DCAT различает прямую ссылку на файл (downloadURL) и адрес доступа (accessURL) — страницу или сервис, через которые данные можно получить.

Доверенная исследовательская среда — защищённая среда, где одобренный исследователь работает с конфиденциальными данными, не скачивая их; результаты проверяют перед выдачей. Пример — SecureLab в UK Data Service.

Доступ по запросу (gated) — набор, описание которого открыто всем, а файлы выдаются после запроса: пользователь сообщает о себе, принимает условия, иногда ждёт одобрения автора. Так устроены закрытые наборы на Hugging Face; для поисковика данных такой набор — набор без файлов.

Единый вход (SSO) — вход в каталог учётной записью внешней системы: ведомства, университета, облачного поставщика или государственной системы идентификации, а не отдельным паролем портала. Проверку личности делает внешняя система по стандарту SAML 2.0 или OpenID Connect. Уволенный сотрудник теряет доступ к каталогу вместе с остальными системами организации.

Журнал изменений — запись о том, кто, когда и что поменял в описании набора, с возможностью посмотреть прежнее состояние. В CKAN — модуль `activity` с разницей между версиями, в научных репозиториях эту роль играют версии набора.

Запись каталога — сведения о том, как набор попал в каталог: когда его внесли и когда правили описание. Отличается от сведений о самих данных: дата правки записи — не дата обновления данных.

Идентификатор объекта — постоянный идентификатор отдельной записи внутри данных: здания, дороги, организации. По нему соединяют разные наборы без сопоставления названий и геометрий. Пример — GERS у Overture Maps. Подробнее — в главе об идентификаторах.

Издатель — организация, которая отвечает за набор по существу и юридически. Может не совпадать с организацией в программе каталога, от имени которой набор заведён.

ИИ-агент — программа на большой языковой модели, которая сама ищет набор, читает описание и забирает данные через API или сервер MCP.

Инвентаризация данных — список того, что организация уже производит или хранит, включая то, что публиковать нельзя. Из этого списка потом выбирают, что выложить, и записывают очередь. Самим каталогом инвентаризация не является.

Инструкция для агента — описание данных для ИИ-агента: определения полей, единицы, период, ограничения, справочники кодов, правила поиска. Её ведут и проверяют, как карточку. Сервер Data360 Всемирного банка отдаёт её ресурсами MCP.

Инфраструктура пространственных данных (ИПД) — общая система правил, каталогов и сервисов, через которую органы власти страны или региона описывают и отдают свои слои. Видимая часть — геопортал и общий каталог; в ЕС такие инфраструктуры построены по директиве INSPIRE.

Карта сайта (sitemap) — файл со списком адресов страниц. По нему поисковик обходит сайт.

Картографический сервер — программа, которая отдаёт слои карты другим программам по протоколам WMS, WFS, ArcGIS REST и подобным. Примеры — GeoServer, MapServer, QGIS Server, ArcGIS Server. Карточек для поиска сам обычно не ведёт, поэтому без каталога рядом слой находит только тот, кто знает адрес.

Карточка набора (dataset card) — описание набора для машинного обучения для человека: состав, способ сбора и разметки, известные перекосы, ограничения и для чего набор не годится. На Hugging Face — файл `README.md` с разделом метаданных в начале. Похожее описание модели называют карточкой модели.

Карточка решения — короткая запись одной развилки при выборе каталога: контекст, варианты, критерии, решение, обоснование, последствия и событие, после которого выбор открывают заново. Пишут на платформу, обязательное расширение и отказ от функции. Заменённую карточку не удаляют, а помечают как заменённую.

Каталог API (портал разработчика) — каталог программных интерфейсов организации или страны: документация, примеры запросов, часто выдача ключей и тестовая консоль. Описывает сервисы, а не наборы; хороший каталог API связывает интерфейс с наборами, которые тот отдаёт.

Каталог данных — сервис, который описывает наборы данных метаданными и помогает их найти, понять и получить.

Каталог каталогов — перечень порталов и репозиториев данных, а не наборов. Помогает понять, где искать данные по теме или стране. Примеры — re3data, dataportals.org, реестр dataportals-registry.

Каталог показателей — каталог официальной статистики, где единица публикации — показатель с разрезами по годам, территориям и другим признакам. Пользователь сам выбирает нужный срез, а не скачивает «файл за год».

Класс программного обеспечения — вид программы, на которой сделан каталог: облачный сервис, готовая платформа на своём сервере, картографический сервер или сервер протокола, сайт на CMS, своя разработка. От класса зависит, кто обновляет программу и сколько стоит уйти; от типа каталога — что в нём публикуют.

Классификатор (справочник, НСИ) — список кодов и названий: регионы, виды деятельности, валюты. На него ссылаются другие наборы, поэтому у классификатора важны версия и дата.

Книга кодов (codebook) — описание файла обследования: переменные, тексты вопросов, коды ответов и их подписи, часто с частотами. Без неё файл микроданных — столбцы чисел без смысла. Машиночитаемый вид — DDI Codebook.

Контролируемый словарь — заранее согласованный список терминов или кодов, из которого выбирают значение поля вместо свободного текста: темы, языки, типы файлов, страны. Тезаурус добавляет к нему синонимы и связи «шире — уже». Хороший словарь даёт каждому термину постоянный адрес и публикуется в SKOS.

Контрольная сумма — короткая строка, вычисленная по содержимому файла алгоритмом вроде MD5 или SHA-256. Меняется при изменении хотя бы одного байта, поэтому по ней проверяют, что скачанный файл цел и что это тот же файл, который опубликован. В DCAT 3 записывается свойством `spdx:checksum`.

Куб данных — таблица показателя со всеми сочетаниями разрезов: годы × регионы × пол и так далее. Из куба вырезают нужный срез, а не скачивают отдельные файлы по годам. Термин из стандартов статистики (SDMX, RDF Data Cube).

Лента изменений — список последних добавленных и изменённых записей каталога в формате RSS или Atom. На неё подписываются программы-сборщики. В CKAN в ленте по умолчанию 20 последних записей, поэтому для полного сбора нужна ещё выборка «изменено после даты».

Лицензия набора — условия, на которых можно использовать конкретный набор. Это отдельный текст от правил всего сайта и от условий доступа к файлу. Открытая лицензия разрешает копировать, менять и использовать данные в том числе в коммерческих целях; указать источник можно требовать, запрещать коммерцию — нет. Часто встречаются CC BY 4.0, CC0 (отказ от прав, насколько это позволяет закон) и ODbL (для баз данных: производную базу отдают на тех же условиях). Одну и ту же лицензию записывают по-разному: код SPDX `CC-BY-4.0`, код CKAN `cc-by`, ссылка словаря ЕС на понятие `CC_BY_4_0`. Фраза «открытая лицензия» и значение «не указана» для программы пусты.

Маркетплейс данных — площадка, где данные продают или выдают по договору: файл, подписку, доступ к интерфейсу или к таблице в облаке. От портала открытых данных отличается договором и условиями на каждую выдачу. Государственные маркетплейсы часто называют биржами данных.

Массовая выгрузка — возможность скачать всю базу или весь набор файлами, без запросов по одному показателю или записи.

Международная хартия открытых данных — шесть принципов публикации государственных данных, согласованных в 2015 году правительствами, общественными организациями и экспертами: открытость по умолчанию, своевременность и полнота, доступность и пригодность к использованию, сопоставимость и совместимость, улучшение управления и вовлечение граждан, инклюзивное развитие и инновации.

Менеджер библиографии — программа, где исследователь хранит свои источники и откуда вставляет ссылки в текст по нужному стилю: Zotero, EndNote, Mendeley, JabRef. Принимает ссылки файлами BibTeX, RIS или CSL-JSON и умеет сохранять их прямо со страницы по метатегам.

Метаданные — данные о данных: название, описание, издатель, лицензия, охват, формат.

Метатеги для библиографии — строки в заголовке HTML-страницы (`citation_title`, `citation_author`, `citation_doi` и другие, а также поля Dublin Core), по которым Zotero и Google Scholar понимают, что перед ними, и сохраняют ссылку одной кнопкой. Версию и тип «набор данных» этими метатегами не передать.

Метрика качества — доля наборов, у которой записаны знаменатель, правило «что считать заполненным» и дата подсчёта. По такой доле видно, что чинить. Число наборов в подвале сайта этой записи не имеет.

Микроданные — записи об отдельных людях, домохозяйствах или организациях из переписей и обследований, в отличие от готовых итоговых таблиц. Перед публикацией их обезличивают, а доступ к файлам часто ограничивают.

Модерация (проверка перед публикацией) — шаг, на котором куратор, модератор или администратор организации смотрит подготовленный набор и выпускает его или возвращает автору. От черновика отличается вторым человеком: черновик автор публикует сам. Модерируют и тексты пользователей — запросы данных, комментарии, обсуждения.

Набор данных — опубликованная совокупность данных с общим описанием.

Наборы данных высокой ценности — шесть групп данных, которые органы власти ЕС по регламенту 2023/138 обязаны отдавать бесплатно, в машиночитаемом виде и через API: геоданные, наблюдение Земли и окружающая среда, метеорология, статистика, компании и их владельцы, мобильность.

Научный репозиторий — каталог, который принимает и надолго хранит данные исследований, выдаёт им постоянные идентификаторы (чаще DOI) и помогает на них ссылаться.

Недокументированный (неявный) API — программный интерфейс, который сайт отдаёт, хотя владелец его не описал: API, из которого страница сама получает данные, или интерфейс платформы по умолчанию вроде `/wp-json/` у WordPress. Полезен, чтобы забрать данные, но меняется без предупреждения.

Обезличивание — обработка данных о людях, после которой по оставшимся полям нельзя узнать конкретного человека. Удалить имя недостаточно: человека часто выдаёт сочетание возраста, адреса и профессии.

Облачный сервис (SaaS) — платформа, которую не ставят на свой сервер, а арендуют у поставщика по подписке. Обновления и хостинг на стороне поставщика, но и данные, и возможности зависят от договора.

Обработка рядом с данными — каталог отдаёт не файл, а выборку или результат расчёта, выполненного на сервере. Примеры — ERDDAP, openEO, Climate Data Store.

Общедоступные и открытые данные — общедоступные можно посмотреть на сайте или карте; открытые можно скачать целиком в машиночитаемом виде, со свободной лицензией и без заявок. Бесплатная выдача по заявлению — ещё не открытые данные.

Общий хаб — чужая площадка, куда автор публикует свой набор вместе с карточкой: Zenodo, Figshare, Hugging Face, Kaggle, Source Cooperative. Хаб данных — другое: несколько каталогов одной организации под общей вывеской. ML-хаб — вид общего хаба, на нём рядом с наборами лежат модели.

Объектное хранилище — облачный или собственный склад файлов, где каждый файл лежит по своему ключу в «бакете» и отдаётся по HTTP: Amazon S3, Azure Blob Storage, Google Cloud Storage и совместимые с S3 системы вроде MinIO и Ceph. Место почти не ограничено, но хранение и скачивание оплачиваются. Большинство крупных каталогов держит загруженные файлы именно там.

Онтология — строгое описание области знания: классы понятий, их свойства и правила, по которым программа может делать выводы. Строже словаря: кроме «шире — уже» задаёт, например, что у гена есть место на хромосоме. Записывается обычно в OWL; каталоги онтологий — OLS, BioPortal, AgroPortal.

Оплата трафика читателем (Requester Pays) — режим облачного бакета, при котором место оплачивает владелец, а скачивание — тот, кто читает файл. Для этого нужна учётная запись облака. Анонимно такой файл не скачать, поэтому для открытых данных режим не подходит.

Ортофотоплан — аэро- или космический снимок, исправленный так, что его можно накладывать на карту и мерить по нему расстояния. Один из самых востребованных слоёв геопорталов.

Отраслевой стандарт данных — общий формат для одной темы, который описывает не файл вообще, а предмет: контракт (OCDS), расписание транспорта (GTFS), наблюдение вида (Darwin Core). Данные в таком формате сразу понимают программы и агрегаторы.

Охраняемые данные — данные государственного сектора, которые нельзя опубликовать открыто: персональные, коммерческая тайна, статистически конфиденциальные. По европейскому Акту об управлении данными их описания собирают в единой информационной точке страны, а сами данные выдают по запросу на условиях, которые сохраняют защиту.

Паспорт набора открытых данных — российское описание набора по методическим рекомендациям версии 4.0 (протокол от 12 декабря 2024 года № 49пр): идентификатор, название, владелец, ссылка на файл, формат и даты. Машиночитаемый вид — файл `meta.csv`, `meta.xml` или `meta.json`. Это не профиль DCAT: международные сборщики и поисковики данных его не читают.

План управления данными — документ к заявке на грант: какие данные появятся в проекте, в каких форматах, где их будут хранить, в каком репозитории и на каких условиях опубликуют. Его требуют Horizon Europe, NIH и многие другие фонды.

Подписанная (временная) ссылка — адрес файла в объектном хранилище с подписью и сроком действия, после которого он перестаёт работать. Порталы выдают её при скачивании, чтобы файл шёл из облака напрямую: у HDX такая ссылка живёт несколько минут. Ссылаться на неё в статье или программе нельзя — только на постоянный адрес файла на портале.

Поисковик данных — сервис, который собирает описания наборов из многих каталогов или со страниц сайтов и даёт по ним один поиск. Своих наборов не публикует, файлы остаются у источника. Бывает общим (Google Dataset Search, Dateno), научным (OpenAIRE, DataCite Commons), отраслевым (NIAID Discovery Portal) и национальным (find.data.gov.scot). Если у каждого собранного набора есть своя страница для обычных поисковиков, его называют агрегатором.

Поисковый индекс — отдельная копия описаний наборов, устроенная для быстрого поиска и фильтров; в каталогах это обычно Solr, Elasticsearch или OpenSearch. Если индекс разошёлся с базой описаний, набор существует, но не находится.

Полнота и детальность метаданных — полнота показывает, какая доля нужных полей описания заполнена у наборов каталога; детальность — насколько подробна сама схема описания. Третье свойство — достоверность: правда ли записано в поле. Значение-заглушка вроде периода «1900–2099» делает поле формально заполненным, но не полным по существу.

Портал открытых данных — каталог, через который орган власти или организация публикует наборы под открытыми лицензиями: реестры, бюджеты, статистику, сведения о транспорте. Открытыми данные делают три условия сразу: машиночитаемый вид, открытая лицензия и доступ без заявок и регистрации.

Посадочная страница (landing page) — страница, на которую ведёт DOI или другой постоянный идентификатор: описание набора, готовая ссылка для цитирования, путь к файлам. Сам идентификатор файлов не хранит, поэтому без живой посадочной страницы он ведёт в никуда.

Постоянный идентификатор (PID) — имя объекта, которое не меняется, и служба (резолвер), которая по этому имени отправляет на текущий адрес объекта. Примеры — DOI, Handle, ARK. Постоянство обеспечивает не технология, а владелец, который обновляет адрес при каждом переезде.

Постраничная выдача (пагинация) — ответ API или выгрузки, разбитый на страницы с определённым числом записей. Программа идёт со страницы на страницу до последней. У многих API есть предел глубины: дальше определённой записи страницы не отдаются.

Пример использования (reuse) — приложение, статья, исследование или визуализация, сделанные на основе набора и привязанные к его карточке на портале. По ним видно, кому набор нужен; образец такого раздела — на data.gouv.fr.

Проверка на входе — каталог не публикует набор, который не прошёл проверку по стандарту или схеме. Примеры — BIDS у OpenNeuro, NWB у DANDI. Подробнее — в главе об уровнях качества.

Производный набор (derived dataset) — запись с DOI для данных, которые пользователь отфильтровал или собрал из нескольких выгрузок, со списком исходных наборов. Так устроено в GBIF.

Пространство данных — договорённость организаций об обмене данными, в том числе не для всех. Описание набора при этом можно прочитать, файл может оставаться закрытым.

Профиль метаданных — уточнение общего словаря под страну или отрасль: какие поля обязательны, какие рекомендуются и из какого справочника берётся значение. Сам словарь при этом не меняется: DCAT-AP и DCAT-US — профили одного DCAT.

Публикатор — человек, который выкладывает файл и заполняет карточку по образцу. Лицензию и содержимое он не назначает: это делают юрист и владелец данных.

Пять звёзд открытых данных — лестница Тима Бернерса-Ли (2010) для одного набора: от файла под открытой лицензией к машиночитаемой таблице, открытому формату, адресам внутри данных и связям с чужими данными. Лестница говорит о формате одного файла.

Разбиение набора (части, splits) — деление набора для машинного обучения на обучающую часть (на ней учат модель), проверочную (на ней подбирают настройки) и тестовую (на ней измеряют результат). Разбиение фиксируют заранее, иначе результаты разных авторов несравнимы.

Разрез (измерение) — признак, по которому делят значения показателя: год, регион, пол, возраст, отрасль. Допустимые значения разреза берутся из справочника с кодами.

Расширение (плагин) — отдельный модуль со своим автором, выпусками и совместимостью с версией ядра. Его ставят рядом с платформой, чтобы получить возможность, которой нет в поставке. Например, ckanext-dcat добавляет CKAN экспорт в DCAT. Настройка или модуль из той же поставки, который только включают, расширением не считается: его выпускают вместе с ядром.

Реестр метаданных — каталог, который хранит не данные, а их описания с версиями и статусами: элементы данных, справочники, структуры таблиц. Стандарт для реестров элементов данных — ISO/IEC 11179, для структур статистики — SDMX.

Роль — набор прав, который выдают человеку в пределах организации, коллекции или сообщества: участник видит закрытые наборы, редактор создаёт и правит, администратор ещё и управляет людьми. Права дают роли, а не человеку, чтобы при смене сотрудника не перебирать отдельные разрешения.

Самоаудит каталога — один проход по своему или чужому каталогу: выгрузка описаний, выборка карточек, подсчёт файлов и страница с тем, что чинить в ближайший месяц. Это проверка конкретного сайта. Рейтинг страны и сертификат архива отвечают на другие вопросы.

Сбор метаданных (харвестинг) — автоматическое периодическое получение метаданных одного каталога другим.

Сервис данных — программный интерфейс, через который данные набора получают по частям: строки таблицы с фильтром, объекты на карте. В DCAT с версии 2 — отдельная сущность со своим адресом, на которую ссылается дистрибутив.

Серия наборов данных — однотипные наборы, которые выходят по очереди: исполнение бюджета за каждый месяц, перепись раз в десять лет. Отдельный класс для серий появился в DCAT 3 (2024); большинство порталов их не знает и заводит каждый выпуск отдельной карточкой.

Синтетические данные — искусственные записи, которые повторяют статистические свойства настоящих, но не описывают реальных людей. Их публикуют вместо закрытых микроданных.

Система координат — правило, по которому числа координат превращаются в точку на Земле. У каждого слоя она своя; указывают её кодом из реестра EPSG, например EPSG:4326 для WGS 84. Слой без кода системы координат нельзя надёжно совместить с другими.

Словарь полей — описание колонок таблицы: название, смысл, единица измерения, допустимые значения. Без него файл открыт формально и непонятен по существу.

Согласование содержимого (content negotiation) — способ получить по одному адресу разные представления: браузер по адресу `https://doi.org/…` получает страницу набора, а программа, которая попросила в заголовке `Accept` формат BibTeX или RIS, — готовую ссылку от DataCite.

Список наборов — страница, таблица или файл README со ссылками на наборы или на чужие каталоги, без поиска и карточек. Нормальный способ опубликовать десяток наборов, если у каждой строки есть лицензия, дата и постоянный адрес.

Стандарт метаданных — общий язык описания набора снаружи: что это, кто издатель, за какой период, где скачать и на каких условиях. Его читают другие каталоги, поисковики и программы. Примеры — DCAT, DataCite, ISO 19115, schema.org. От стандарта данных отличается тем, что не описывает колонки внутри файла.

Статический сайт — сайт из готовых HTML-файлов. Генератор собирает их из исходников, хостинг только отдаёт файлы: отдельной базы и сервера приложений нет.

Статья о данных (data paper) — рецензируемая статья, которая описывает набор данных и ссылается на него в репозитории, без гипотез и выводов исследования.

Страница-надгробие (tombstone page) — страница, которую оставляют на месте снятого набора с DOI: ссылка для цитирования, сам DOI и объяснение, почему данных больше нет. Нужна потому, что DOI нельзя удалить.

Ступенчатый доступ — доступ к набору по ступеням: описание, переменные и сводная статистика, выборка-образец, все данные. Каждую ступень открывают отдельно: всем, отдельным людям или после условий. Пример — Redivis.

Тема и тег — тема берётся из общего справочника (темы Бюро публикаций ЕС, тематические категории ISO 19115), поэтому по ней наборы разных порталов складываются вместе. Тег — свободное слово публикатора; для программы «транспорт» и «Транспорт» — разные теги.

Техническое задание (ТЗ) — документ, по которому подрядчик делает или настраивает систему: что она должна уметь, на каких условиях и как проверят результат. Для каталога данных это прежде всего список требований.

Тип каталога — класс каталога по тому, что в нём публикуют и что получает свою страницу: набор, слой карты, версия набора с DOI, показатель, обследование и так далее. Задаёт стандарт описания и круг подходящих платформ.

Типовой портал — одна и та же сборка портала, которую государство или поставщик раздаёт десяткам и сотням органов власти: у каждого своя витрина, но программа, поля описания и выгрузка общие. Примеры — GD Catalog в Таиланде, BODIK ODCS в Японии, Municipium в Италии.

Требование к каталогу — функция, которую выбрали для своего каталога, записанная вместе с источником потребности, уровнем обязательности (обязательно, желательно, не нужно), обоснованием, способом проверки и условием пересмотра. Хорошее требование говорит, что нужно, а не как это сделать: «национальный портал забирает описания в DCAT-AP», а не «поставить ckanext-dcat».

Универсальный репозиторий (generalist) — научный репозиторий, который принимает данные любой дисциплины от любого автора: Zenodo, Figshare, Harvard Dataverse, Dryad. Фонды советуют его, когда у дисциплины нет своего предметного репозитория.

Утечка тестовых данных (contamination) — попадание тестовой части бенчмарка в обучающие данные модели, чаще всего вместе с текстами из интернета. Результат модели на таком бенчмарке завышен. Защита — скрытые ответы тестовой части и особые метки в файлах, по которым их выбрасывают из обучающих наборов.

Файл изменений (дельта) — файл, в котором только то, что изменилось с прошлой полной выгрузки: новые, изменённые и удалённые записи. Позволяет не скачивать большую базу целиком при каждом обновлении.

Файл общего пользования (public use file) — обезличенный файл микроданных, который выдают после регистрации и согласия с условиями, без рассмотрения заявки. Подробнее и строже — лицензированный файл по одобренной заявке, ещё строже — работа только в доверенной исследовательской среде.

Форк — копия чужого репозитория, которую ведёт другая команда. Копии расходятся: ставят ту, которая проверяется на вашей версии ядра, и записывают её адрес. Правка, которую оставили только у себя и не отдали обратно, тоже становится форком, и сопровождающий тогда — вы.

Хаб данных — несколько каталогов разных типов под одной вывеской и с общим поиском, например портал набора, каталог показателей и каталог микроданных одной международной организации. Не отдельный тип: каждый каталог хаба оценивают по своему типу.

Хранилище таблиц — база данных каталога, куда загружены строки таблиц из файлов, чтобы показывать их в браузере, фильтровать и отдавать частями через API. В CKAN — DataStore на PostgreSQL, у облачных сервисов вроде Huwise и Socrata набор и есть таблица в их базе. Рядом всё равно нужна выгрузка целиком.

Цитирование данных — ссылка на набор данных в статье, отчёте или другом наборе, оформленная как ссылка на публикацию: автор, год, название, версия, издатель, идентификатор. Отличается от требования лицензии указать автора: лицензия задаёт юридическое условие, ссылка даёт научное признание и позволяет найти именно те данные, на которых сделан вывод.

Шлюз API (API gateway) — программа, которая стоит перед интерфейсами организации: проверяет ключи, считает запросы, ограничивает нагрузку. В комплекте обычно идёт портал разработчика, который и становится каталогом API. Примеры — Azure API Management, WSO2 API Manager, Gravitee.

Элемент данных — описание одного поля, общее для многих систем: понятие (объект и свойство, например «человек — пол»), допустимые значения, определение, хранитель, версия и статус. Так поля описывают реестры по ISO/IEC 11179.

Эмбарго — срок, до которого описание набора открыто, а сами данные закрыты; после даты данные открываются. Обычно в научных репозиториях, пока не вышла статья.

Ядро платформы — возможности, которые входят в поставку программы. Сюда же относят модуль из той же поставки, который только включают в настройках: у него тот же автор и тот же цикл обновлений. Отдельный репозиторий со своей совместимостью — уже расширение.

API — вход, по которому программа задаёт каталогу вопрос и получает ответ. Ответ чаще всего приходит в JSON.

ArcGIS REST — программный интерфейс серверов и облака ArcGIS компании Esri. Не стандарт, но из-за распространённости ArcGIS его читают программы GIS и многие сборщики наравне со стандартами OGC.

ARK — постоянный идентификатор вида `ark:12148/…`, который выдают библиотеки, архивы и музеи. Номер организации бесплатен, разрешать ARK можно на своём сервере или через общий резолвер n2t.net.

AsyncAPI — формат описания интерфейсов, которые не отвечают на запрос, а присылают поток сообщений: события о задержках транспорта, показания датчиков. Делает для потоков то же, что OpenAPI для обычных запросов.

BagIt — стандарт упаковки файлов для архивного хранения и передачи: данные лежат в папке рядом с манифестом контрольных сумм, так что по пакету видно, все ли файлы на месте и не изменились ли они.

BibTeX, RIS, CSL — форматы, в которых каталог отдаёт ссылку на набор менеджеру библиографии. BibTeX — для LaTeX; отдельного типа для данных в нём нет, его добавил biblatex (`@dataset`). RIS — для Zotero, EndNote и подобных, тип данных в нём `DATA`. CSL — стили, по которым ссылку оформляют под требования журнала или ГОСТ, и формат CSL-JSON для передачи самой ссылки.

BIDS — стандарт организации данных нейровизуализации: порядок папок, имён файлов и описаний. По нему OpenNeuro проверяет наборы при загрузке.

CDIF — рекомендации CODATA, какие уже существующие стандарты сочетать, когда данные переходят границу научной дисциплины. Для нахождения советует schema.org и DCAT, а не новый словарь каталога. Версия 1.1 вышла в июне 2026 года.

CITATION.cff — текстовый файл в корне репозитория, где написано, как ссылаться на данные или программу. GitHub показывает по нему кнопку «Cite this repository». Zenodo берёт из него авторов и название при архивации выпуска GitHub, если рядом нет `.zenodo.json`; если этот

файл есть, Zenodo читает только его. Для данных в `CITATION.cff` нужна строка `type: dataset`. Тип записи в Zenodo задают отдельно, полем `upload_type`.

Cloud Optimized GeoTIFF (COG) — файл GeoTIFF, устроенный так, что программа читает нужный фрагмент снимка по сети и не скачивает файл целиком.

CoreTrustSeal — знак того, что репозиторий умеет хранить данные долго, а не только показать их на сайте.

CRIS — система учёта научной работы организации: публикации, проекты, гранты, сотрудники и в том числе наборы данных. Пример — Elsevier Pure. Наборы в ней часто служат скрытым каталогом данных университета.

Croissant — машиночитаемое описание набора для машинного обучения поверх schema.org: какие файлы, какие поля, как набор делится на обучающую и тестовую части. Разработан MLCommons. Версия 1.1 (январь 2026 года) добавила происхождение данных и условия использования; расширение RAI описывает перекосы, ограничения и назначение набора. Площадки в 2026 году ещё отдают описания по версии 1.0.

Crossref — агентство регистрации DOI для статей, книг, препринтов и других публикаций, работает по законам США. Для наборов данных обычно выбирают DataCite, но InvenioRDM и Dataverse умеют регистрировать DOI и через Crossref.

CSV on the Web (CSVW) — рекомендация W3C 2015 года: словарь полей таблицы CSV в отдельном файле JSON — названия колонок, типы, единицы, связи с другими таблицами. Google читает такой словарь из разметки страницы набора в экспериментальном режиме и предупреждает, что проверка расширенных результатов на этот блок ругается.

CSW — протокол, которым программа запрашивает каталог слоёв геопортала.

Darwin Core — стандарт описания наблюдений и образцов живых организмов: какой вид, где и когда. В нём данные собирает GBIF.

DataCite — некоммерческая организация, которая регистрирует DOI для данных, и одноимённая схема описания, которую заполняют при регистрации.

Dataspace Protocol — протокол обмена в пространствах данных, который ведёт Eclipse Foundation (версия 2025-1). Участник публикует каталог наборов в DCAT с условиями в ODRL, а протокол описывает переговоры о договоре и передачу данных.

DCAT — словарь W3C для описания каталогов, наборов, способов получения и сервисов данных. Версия 3 — рекомендация от 22 августа 2024 года: в ней появились серии наборов и свойства версии.

DCAT-AP — профиль DCAT для порталов Европы: какой набор полей описания они отдают друг другу. Действующая рекомендация — 3.0.1 от 27 октября 2025 года. Черновик 3.0.2 от 17 июня 2026 года рекомендацией ещё не стал.

DCAT-US — профиль DCAT для федеральных ведомств США. Ведомство публикует файл `data.json` со списком своих наборов, а data.gov его собирает. Версия 3.0 опубликована 1 мая 2026 года и построена на DCAT 3. Версию 1.1 data.gov принимает на переходный период.

DDI — стандарт описания обследований и микроданных: кто спрашивал, кого, как строили выборку и что означают переменные. Семья из нескольких форматов: DDI Codebook описывает одно исследование (основная версия 2.5, новая 2.6 — 2026), DDI Lifecycle — весь цикл с вопросниками и волнами (3.3), DDI-CDI — данные любой области (1.0, 2025). Общая часть с 2026 года — стандарт ISO/PAS 25955.

DOI — постоянный идентификатор цифрового объекта вида `10.5281/zenodo.16546069`: префикс организации и суффикс, который она придумывает сама. Ссылка `https://doi.org/...` перенаправляет на посадочную страницу. Для данных DOI обычно регистрируют в DataCite, для статей — в Crossref; технически это Handle с префиксом `10.`. Удалить DOI нельзя. Общий (родительский) DOI объединяет все версии набора и ведёт на последнюю.

Dolt — база данных с языком SQL, совместимая с MySQL, с версиями как в Git: клон, ветки, слияние, разница по ячейкам. Публичные базы хранит DoltHub. Подробнее — в главе о Git-репозиториях.

Dublin Core — короткий набор полей описания: название, автор, дата, тема. Его часто отдают научные репозитории.

FAIR — принципы находимости, доступности, совместимости и повторного использования данных (2016). FAIR не значит «открыто»: набор с закрытыми файлами тоже может им соответствовать, если у него открытое описание, постоянный идентификатор и понятный порядок доступа.

FAIR Data Point — небольшой сервис, через который организация рассказывает в DCAT, какие данные у неё есть. Сами данные при этом остаются у владельца.

FAIR Signposting — способ подсказать программе главное о наборе без разбора HTML: страница набора в заголовке ответа или в отдельном файле (linkset) перечисляет ссылки на описание, файлы, лицензию и идентификаторы авторов. Есть в ядре Dataverse, InvenioRDM и DSpace.

Frictionless Data Package — стандарт описания набора данных файлом рядом с данными. Опубликованный описатель называется `datapackage.json`. Текущая версия стандарта — 2.0.

GeoDCAT-AP — профиль DCAT-AP для геоданных: как перевести карточку слоя по ISO 19115 в описание DCAT, понятное порталам открытых данных. По нему data.europa.eu читает национальные геокаталоги.

GeoPackage — открытый формат OGC для геоданных: один файл на основе базы SQLite, в котором лежат векторные слои, растры и их описание. Открывается в QGIS и других программах без сервера.

Git LFS — дополнение к Git для больших файлов: сами файлы хранятся отдельно, а в истории репозитория остаются ссылки на них. У площадки на это свои пределы размера файла и квота на хранение и скачивание.

GTFS — формат расписаний общественного транспорта: остановки, маршруты, рейсы. Дополнение GTFS Realtime передаёт положение машин и задержки в реальном времени.

Handle — система постоянных идентификаторов, на которой построены и DOI. Префикс получают в Handle.Net Registry за ежегодный сбор и держат свой сервер Handle. DSpace выдаёт Handle по умолчанию, но с префиксом-заглушкой `123456789`, который вне сайта ничего не значит, пока организация не зарегистрирует свой.

HealthDCAT-AP — профиль DCAT-AP для описаний медицинских данных: орган доступа, категория данных, число людей в наборе, возрастной охват, правовое основание. Его готовит Еврокомиссия для каталогов, которые требует регламент о Европейском пространстве медицинских данных (EHDS); в 2026 году это проект спецификации.

IATI — стандарт раскрытия данных о помощи развитию: кто финансирует проект, сколько, где и с каким результатом. Публикуют международные организации, доноры и фонды.

INSPIRE — правила ЕС для пространственных данных: как описывать слои и как отдавать их каталогом. Действуют с 2007 года, с декабря 2025 года директиву упрощают; отдельный геопортал INSPIRE закрыт 1 июля 2026 года, его роль перешла к data.europa.eu.

ISO 19110 — стандарт каталога объектов для геоданных: какие типы объектов есть в слое и какие у них атрибуты. По сути — словарь полей слоя карты. Поддерживается в ядре GeoNetwork.

ISO 19115 — стандарт описания слоя карты: что изображено, за какой период, кто автор, в какой системе координат. Карточками обмениваются в XML по ISO 19139, новая редакция записывается по ISO 19115-3.

ISO/IEC 11179 — стандарт реестров метаданных: как описать отдельный элемент данных — его определение, допустимые значения, владельца и статус.

JSON — текстовый формат, которым программы обмениваются данными. Человеку он читается хуже таблицы, программе — лучше, чем PDF.

JSON Lines — файл, в котором каждая строка — отдельная запись JSON. Его читают построчно, не загружая целиком в память, поэтому он подходит для больших выгрузок со вложенными полями.

JSON-stat — лёгкий формат статистической таблицы в JSON: значения, разрезы и подписи в одном ответе. Версия 2.0 — формат по умолчанию у API Евростата и PxWebApi 2.

llms.txt — короткий текстовый файл о сайте для языковой модели: что здесь опубликовано и куда смотреть.

MCP — протокол Model Context Protocol, по которому каталог отдаёт ИИ-агенту готовый список запросов, справочники и инструкции, чтобы тот не угадывал адреса. Сервер MCP работает поверх API каталога. Хороший сервер — только на чтение, без ключа для открытых наборов и со ссылкой на набор в ответе. Подробнее — в главе об ИИ.

ML-каталог — каталог наборов данных, а часто и моделей, для обучения и проверки систем машинного обучения. Набор в нём разбит на части, загружается из кода и описан карточкой с перекосами и ограничениями.

ML-хаб — общий хаб, куда любой загружает наборы и модели: Hugging Face, Kaggle, ModelScope. Хаб данных — другое: несколько каталогов одной организации под общей вывеской.

MQA — оценка карточек, которые data.europa.eu забрал у портала: заполнены ли поля DCAT-AP и отвечают ли ссылки. Балл каталога — от 0 до 7,5 и среднее по тем описаниям, которые оценка посчитала. Портал, который в европейский каталог ничего не отдаёт, балла не получает.

NetCDF — формат многомерных массивов: переменные на сетке по времени и пространству. С соглашениями CF в нём хранят климатические, океанографические и атмосферные данные.

OAI-PMH — протокол, которым научные репозитории отдают описания записей сборщику. Передаёт только описания, без списка файлов, поэтому собранная по нему карточка часто не ведёт к самим данным.

OCDS — стандарт открытых данных о закупках: весь путь контракта от плана и торгов до исполнения.

ODIN — оценка официальной статистики стран, которую ведёт Open Data Watch. Смотрит сайт статистической службы и страницы, на которые она ссылается: есть ли данные по 22 темам и открыты ли они. Портал открытых данных страны в эту оценку может не входить.

ODRL — язык W3C для машиночитаемых условий использования: что разрешено, что запрещено, что обязательно сделать. Применяется в пространствах данных.

OGC API — семейство современных веб-интерфейсов к геоданным от консорциума OGC: Features для объектов, Records для каталога слоёв и другие. Постепенно заменяет WFS и CSW.

Open Data Maturity — ежегодная оценка европейских стран с 2015 года: политика, портал, качество метаданных и влияние, поровну. Это опрос национальных представителей, не обход файлов. В отчёте 2025 года участвовали 36 стран.

OpenAIRE — европейская инфраструктура открытой науки. Собирает описания публикаций, наборов данных и программ из репозиториев по OAI-PMH и из DataCite в общий граф научных результатов и даёт поиск по нему (OpenAIRE Explore). Сервис OpenAIRE CONNECT строит над этим графом отдельные витрины для научных сообществ.

OpenAPI — формат описания программного интерфейса: какие запросы он принимает и что отвечает. По описанию строят документацию и проверяют запросы.

openEO — открытый протокол обработки спутниковых данных на сервере: пользователь отправляет цепочку операций, сервер возвращает результат.

OpenID Connect — стандарт проверки личности поверх протокола OAuth 2.0: каталог отправляет пользователя войти во внешнюю систему и получает подтверждение, кто он. На нём работают Keycloak, облачные системы учётных записей, французский ProConnect.

OpenSharing (Delta Sharing) — открытый протокол, по которому облачная площадка даёт покупателю доступ к таблицам поставщика без копирования данных. Создан Databricks как Delta Sharing в 2021 году, в июне 2026 года стал OpenSharing и перешёл в Linux Foundation.

ORCID — постоянный идентификатор исследователя. По нему наборы и статьи связываются с автором, даже если его имя пишут по-разному.

OWL — язык онтологий W3C: классы, свойства и правила вывода поверх RDF. В нём записаны многие онтологии наук о жизни и сельского хозяйства.

Parquet — колоночный формат таблиц: хранит типы колонок и сжимает данные по колонкам. Файл в несколько раз меньше CSV, а аналитические программы читают его быстрее.

PC-Axis (PX) — файловый формат статистических таблиц со встроенным описанием разрезов и единиц. Родной формат PxWeb; его отдают и многие порталы открытых данных.

Piveau — свободная платформа каталогов, построенная вокруг DCAT и сбора чужих каталогов. На ней работает data.europa.eu.

PURL и w3id.org — бесплатные службы постоянных адресов: адрес на purl.org или w3id.org перенаправляет туда, где ресурс лежит сейчас, и при переезде меняется только правило перенаправления. purl.org ведёт Интернет-архив, w3id.org — сообщество W3C, правила в нём добавляют запросом в репозиторий на GitHub.

RDF — общая модель записи сведений в виде утверждений «объект — свойство — значение». На ней построены DCAT, SKOS и другие словари связанных данных.

re3data — международный реестр научных репозиториев: тематика, условия доступа, сертификаты.

RFC 9727 (api-catalog) — стандарт IETF 2025 года: по адресу `/.well-known/api-catalog` сайт отдаёт программе список своих интерфейсов. В 2026 году встречается редко.

RO-Crate — способ упаковать файлы исследования вместе с описанием: рядом с данными лежит `ro-crate-metadata.json` на основе schema.org. Так сдают и передают объект исследования. Каталог страны этим файлом не собирают. Рекомендация 1.3 — от 22 июня 2026 года.

ROR — открытый реестр научных организаций с постоянными идентификаторами. В описании набора ссылка на ROR указывает организацию автора или издателя однозначно.

SAML 2.0 — стандарт OASIS 2005 года для единого входа: система учётных записей организации передаёт каталогу подписанное подтверждение, кто вошёл и какие у него признаки. Основа университетских федераций входа и Shibboleth.

schema.org/Dataset — разметка страницы набора данных, по которой его находят поисковики. Google требует у описания от 50 до 5000 знаков.

SDMX — стандарт обмена статистикой: не файл за один год, а ряд с разрезами. Включает модель описания таблиц (разрезы, справочники, атрибуты), форматы SDMX-ML, SDMX-JSON и SDMX-CSV и правила API. Его ведут семь международных организаций, среди них Евростат, МВФ и ОЭСР; действующая версия — 3.1 (2025), но большинство сервисов работает по 2.1.

SKOS — способ описать словарь терминов и связи между ними (рекомендация W3C 2009 года): у каждого понятия адрес, названия на разных языках, код, связи «шире — уже» и соответствия понятиям других словарей.

SPDX — короткие коды лицензий, чтобы программа узнала лицензию без разбора юридического текста.

STAC — протокол описания пространственных файлов: где лежит файл, к какой территории и дате относится, как его скачать. Появился для спутниковых снимков, теперь им описывают и другие геоданные. Это не программа: STAC отдают отдельные серверы, статические наборы JSON-файлов и модули универсальных платформ вроде GeoServer.

StatDCAT-AP — дополнение к DCAT-AP для статистических таблиц: разрезы, атрибуты, ссылка на структуру SDMX. Нужно, чтобы таблицы статистической службы появились на портале открытых данных. Действующая версия — 1.0.1 (2019). Версия 3.0.0 с 10 июня 2026 года — рабочий черновик, приведённый к DCAT-AP 3.0.1.

Table Schema — стандарт семейства Frictionless для словаря полей таблицы: названия колонок, типы, допустимые значения.

VIVO — свободная программа для профилей учёных организации: люди, публикации, проекты, иногда наборы данных. Каталогом данных в строгом смысле не является, но в реестрах её установки встречаются среди поисковиков.

WARC и WACZ — форматы веб-архива. WARC хранит снимки страниц вместе с ответами сервера, его читают Интернет-архив и национальные библиотеки. WACZ — тот же архив, упакованный в ZIP вместе с индексом, его можно открыть в браузере без сервера.

WMS и WFS — протоколы OGC для геоданных. WMS отдаёт картинку слоя для просмотра, WFS — сами объекты, которые можно скачать и обработать. WMTS отдаёт карту заранее нарезанными плитками, поэтому она быстрее WMS, но тоже только для просмотра.

XBRL — язык финансовой отчётности, в котором каждая строка отчёта помечена машиночитаемым кодом. В нём компании сдают отчётность регуляторам во многих странах.

XKOS — расширение SKOS для статистических классификаций от DDI Alliance: уровни классификации, таблицы соответствия между редакциями, пояснения «включает» и «не включает».

Zarr — формат многомерного массива из отдельных кусков. Программа читает нужный кусок по сети. Так публикуют климатические поля, результаты моделей и другие массивы, которые не влезают в один файл.

Zenodo — общий научный репозиторий при ЦЕРН. Принимает данные любой дисциплины, выдаёт DOI и умеет автоматически архивировать выпуски репозиториев GitHub.

Как работать с реестром

Цифры книги получены из выгрузки dataportals-registry в DuckDB (`data/datasets/datasets.duckdb`), выпуск 1.27.0 от 15 октября 2026 года, запросы от 15 октября 2026 года. Таблица `catalogs` — одна строка на каталог, таблица `software` — одна строка на платформу. Все цифры книги рассчитаны ровно по этому выпуску. Имена типов ниже — как в схеме реестра, не как заголовки глав.

Как получить выгрузку

Файл базы лежит в самом репозитории реестра, в папке `data/datasets`. Рядом те же записи в Parquet и JSONL: `full.parquet` и `full.jsonl`. Реестр обновляют каждые несколько дней, поэтому для проверки цифр книги нужен выпуск, по которому они посчитаны, а не последнее состояние репозитория:

```
git clone https://github.com/datenoio/dataportals-registry.git
cd dataportals-registry
git checkout v1.27.0
duckdb -readonly data/datasets/datasets.duckdb
```

Первым запросом сверьте общее число записей с итогом в главе «Каталоги в цифрах»: в выпуске 1.27.0 он даст 43 614.

```
SELECT count(*) FROM catalogs;
```

Если числа не совпали, у вас другая выгрузка, и остальные запросы тоже дадут другие цифры.

Каталоги по типам

```
SELECT catalog_type, count(*) AS n
FROM catalogs
GROUP BY 1
ORDER BY 2 DESC;
```

Платформы внутри каждого типа

```
SELECT catalog_type, sid, n
FROM (
  SELECT catalog_type, software.id AS sid, count(*) AS n,
         row_number() OVER (PARTITION BY catalog_type ORDER BY count(*) DESC) AS rn
  FROM catalogs
  GROUP BY 1, 2
)
WHERE rn <= 6
ORDER BY catalog_type, n DESC;
```

Доли самописных решений, API и закрытых каталогов по типам

```
SELECT catalog_type,
       round(100.0 * count(*) FILTER (WHERE software.id = 'custom') / count(*), 0) AS
pct_custom,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       round(100.0 * count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) / count(*),
1) AS pct_dead
FROM catalogs
GROUP BY 1
ORDER BY count(*) DESC;
```

Владельцы

```
SELECT owner.type, count(*) AS n
FROM catalogs
GROUP BY 1
ORDER BY 2 DESC;
```

Распространённость интерфейсов

```
SELECT e.type, count(*) AS n
FROM (SELECT unnest(endpoints) AS e FROM catalogs)
GROUP BY 1
ORDER BY 2 DESC
LIMIT 30;
```

Лицензии, национальные каталоги, DOI

Около 420 каталогов с текстом или ссылкой на лицензию, около 2 тысяч с DOI и 613 национальных — этот запрос. Глава о лицензиях.

```
SELECT
  count(*) FILTER (WHERE rights.license_id IS NOT NULL
                      OR rights.license_url IS NOT NULL
                      OR rights.license_name IS NOT NULL) AS with_license,
  count(*) FILTER (WHERE properties.is_national) AS n_national,
  count(*) FILTER (WHERE properties.has_doi) AS with_doi,
  count(*) AS total
FROM catalogs;
```

18 тысяч с определённой моделью лицензии и 5,6 тысячи с явной пометкой «неизвестно» — строки следующего запроса, не сумма с 420. Пустое значение — записи, которые глава не относит ни к одному варианту. Срезы не складываются в общее число каталогов: текст лицензии и модель лицензии — разные поля.

```
SELECT rights.rights_type AS model, count(*) AS n
FROM catalogs
GROUP BY 1
ORDER BY n DESC;
```

Ограниченный доступ, около 1,1 тысячи в той же главе, — значение `restricted` в списке `access_mode`.

```
SELECT mode, count(*) AS n
FROM (SELECT unnest(access_mode) AS mode FROM catalogs)
GROUP BY 1
ORDER BY n DESC;
```

Статусы и программный интерфейс

868 недействующих и 98 выведенных из эксплуатации — строки `inactive` и `deprecated`. 65% и 34% с интерфейсом и без него — доли `api`.

```
SELECT status, count(*) AS n
FROM catalogs
GROUP BY 1
ORDER BY n DESC;
```

```
SELECT
  count(*) FILTER (WHERE api) AS with_api,
  count(*) FILTER (WHERE NOT coalesce(api, false)) AS without_api,
  count(*) AS total
FROM catalogs;
```

Доля интерфейса у списков наборов, около 14%, — тип `Datasets list`.

```
SELECT
  count(*) AS n,
  round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api
FROM catalogs
WHERE catalog_type = 'Datasets list';
```

Карта сайта, около 7 тысяч, — среди типов в запросе «Распространённость интерфейсов». Если нужно только её:

```
SELECT e.type, count(*) AS n
FROM (SELECT unnest(endpoints) AS e FROM catalogs)
WHERE lower(e.type) LIKE '%sitemap%'
GROUP BY 1
ORDER BY n DESC;
```

Владельцы, страны, состав

Местные органы, около 17,6 тысячи, и геопорталы среди них, около 14,4 тысячи:

```
SELECT catalog_type, count(*) AS n
FROM catalogs
WHERE owner.type = 'Local government'
GROUP BY 1
ORDER BY n DESC;
```

Страны в главе о цифрах. Каталог с несколькими странами попадёт в каждую. Записи без одной страны — отдельная строка с пустым или неоднозначным покрытием; в главе их около 1,8 тысячи.

```
SELECT country_id, count(*) AS n
FROM (
  SELECT unnest(list_transform(coverage, x -> x.location.country.id)) AS country_id
  FROM catalogs
)
GROUP BY 1
ORDER BY n DESC
LIMIT 15;
```

Состав содержимого в главе «Что публикуют»: один каталог может дать несколько строк.

```
SELECT t AS content_type, count(*) AS n
FROM (SELECT unnest(content_types) AS t FROM catalogs)
GROUP BY 1
ORDER BY n DESC;
```

Классификация: владельцы, доступ и классы ПО по типам

Глава «Как классифицировать каталоги». Владельцы внутри каждого типа — например, 58% геопорталов у местных органов власти и 85% научных репозиториев у университетов и научных организаций:

```
SELECT catalog_type, owner.type AS owner_type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (PARTITION BY catalog_type), 0) AS pct
FROM catalogs
GROUP BY 1, 2
ORDER BY catalog_type, n DESC;
```

Доля каталогов с ограниченным доступом хотя бы к части данных — 91% у маркетплейсов, 49% у каталогов микроданных:

```
SELECT catalog_type,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) AS restricted,
       count(*) AS n,
       round(100.0 * restricted / n, 0) AS pct
FROM catalogs
GROUP BY 1
ORDER BY pct DESC;
```

Классы ПО — поле `subtype` в таблице `software`. В главе они сведены в пять групп: облачный сервис (`managed_saas_service`), готовая платформа (подтипы `*_platform`), сервер или отраслевая система (`geospatial_service_middleware`, `protocol_or_api_server`, `domain_data_infrastructure`), сайт на CMS (`cms_or_app_framework`) и своя разработка (`custom`).

```
SELECT CASE WHEN c.software.id = 'custom' THEN 'custom' ELSE s.subtype END AS class,
       count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (), 1) AS pct
FROM catalogs c
```

```
LEFT JOIN software s ON s.id = c.software.id
GROUP BY 1
ORDER BY n DESC;
```

Платформы, которые книга называет по имени

Идентификатор — `software.id`. Около 1 900 CKAN считаются только среди порталов открытых данных, не среди всех записей с этим id.

```
SELECT software.id AS sid, catalog_type, count(*) AS n
FROM catalogs
WHERE software.id IN (
  'ckan', 'dkan', 'arcgishub', 'arcgisserver', 'geoserver', 'emapa',
  'geonetwork', 'geonode', 'jkan', 'opensdg', 'datasette',
  'opendatasoft', 'socrata', 'pxweb', 'nada', 'dspace',
  'dataverse', 'inveniordm', 'invenio', 'custom'
)
GROUP BY 1, 2
ORDER BY n DESC;
```

Порталы открытых данных

Глава «Порталы открытых данных». Владельцы — 48% у местных органов власти, 19% у центральных:

```
SELECT owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (), 0) AS pct
FROM catalogs
WHERE catalog_type = 'Open data portal'
GROUP BY 1
ORDER BY n DESC;
```

Страны — Россия 572, США 565, Испания 485. Каталог с несколькими странами попадёт в каждую:

```
SELECT country_id, count(*) AS n
FROM (
  SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id
  FROM catalogs
  WHERE catalog_type = 'Open data portal'
)
GROUP BY 1
ORDER BY n DESC
LIMIT 15;
```

Платформы с долей API и недействующих — CKAN 1906 и 95% с API, своя разработка 1144 и 19%:

```
SELECT software.id AS sid, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       round(100.0 * count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) / count(*),
1) AS pct_dead
FROM catalogs
WHERE catalog_type = 'Open data portal'
GROUP BY 1
```

```
ORDER BY n DESC
LIMIT 25;
```

Платформы у центральных и у местных органов власти — CKAN 420 и 707, Gosweb 421 только у местных:

```
SELECT owner.type, software.id AS sid, count(*) AS n
FROM catalogs
WHERE catalog_type = 'Open data portal'
  AND owner.type IN ('Central government', 'Local government')
GROUP BY 1, 2
QUALIFY row_number() OVER (PARTITION BY owner.type ORDER BY count(*) DESC) <= 10
ORDER BY 1, n DESC;
```

Три главные платформы в каждой стране — так видно, что в России счёт дают разделы на Gosweb, в Италии — Municipium, в Японии — BODIK ODCS:

```
SELECT country_id, sid, n
FROM (
  SELECT country_id, sid, count(*) AS n,
         row_number() OVER (PARTITION BY country_id ORDER BY count(*) DESC) AS rn
  FROM (
    SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id,
           software.id AS sid
    FROM catalogs
    WHERE catalog_type = 'Open data portal'
  )
  GROUP BY 1, 2
)
WHERE rn <= 3 AND country_id IN ('RU', 'US', 'ES', 'FR', 'IT', 'TH', 'ID', 'JP', 'CN')
ORDER BY country_id, n DESC;
```

Геопорталы

Глава «Геопорталы». Владельцы — 58% у местных органов власти. В главе центральные органы сложены с федеральными (`Federal government`), а «Прочие» — это `Other` и `Community`:

```
SELECT owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (), 0) AS pct
FROM catalogs
WHERE catalog_type = 'Geoportal'
GROUP BY 1
ORDER BY n DESC;
```

Страны — США 5050, Польша 2303, Чехия 1408. Три главные платформы в каждой стране — так видны местные поставщики (GISPLAN в Чехии, iObčina в Словении, ⊠⊠⊠⊠⊠ в Японии):

```
SELECT country_id, sid, n
FROM (
  SELECT country_id, sid, count(*) AS n,
         row_number() OVER (PARTITION BY country_id ORDER BY count(*) DESC) AS rn
  FROM (
    SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id,
           software.id AS sid
    FROM catalogs
```

```
    WHERE catalog_type = 'Geoportal'
  )
  GROUP BY 1, 2
)
WHERE rn <= 3
ORDER BY country_id, n DESC;
```

Платформы с долей API и недействующих — ArcGIS Server 4674, своя разработка 1043 и 11% с API, GeoNode 11,2% недействующих. Российские геопорталы (286) считаются по стране владельца: добавьте `AND owner.location.country.id = 'RU'`.

```
SELECT software.id AS sid, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       round(100.0 * count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) / count(*),
1) AS pct_dead
FROM catalogs
WHERE catalog_type = 'Geoportal'
GROUP BY 1
ORDER BY n DESC
LIMIT 40;
```

Облачные сервисы местных поставщиков без ArcGIS Hub и e-mapa — 132 сервиса, около 5,3 тысячи геопорталов, API у 21%; у 104 сервисов API не отмечен ни на одном сайте:

```
WITH v AS (
  SELECT c.software.id AS sid, count(*) AS n, count(*) FILTER (WHERE c.api) AS with_api
  FROM catalogs c
  JOIN software s ON s.id = c.software.id
  WHERE c.catalog_type = 'Geoportal'
    AND s.subtype = 'managed_saas_service'
    AND c.software.id NOT IN ('arcgishub', 'emapa')
  GROUP BY 1
)
SELECT count(*) AS vendors, sum(n) AS sites,
       round(100.0 * sum(with_api) / sum(n), 0) AS pct_api,
       count(*) FILTER (WHERE with_api = 0) AS vendors_without_api
FROM v;
```

Интерфейсы каталога у геопорталов — около 3,2 тысячи выгрузок DCAT-US и DCAT-AP у сайтов ArcGIS Hub, CSW у GeoNetwork и GeoNode, около 140 серверов STAC:

```
SELECT c.software.id AS sid, e.type, count(DISTINCT c.id) AS n
FROM (SELECT id, software, unnest(endpoints) AS e FROM catalogs WHERE catalog_type =
'Geoportal') c
WHERE e.type IN ('dcatus11', 'dcatap201', 'ogcrecordsapi', 'csw202', 'stacserverapi',
'oaipmh20')
GROUP BY 1, 2
HAVING n > 20
ORDER BY 2, n DESC;
```

Научные репозитории

Глава «Научные репозитории данных». Всего 8404, недействующих и выведенных 119, с ограниченным доступом 527:

```
SELECT count(*) AS n,
```

```
       count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) AS restricted
FROM catalogs
WHERE catalog_type = 'Scientific data repository';
```

Владельцы — 85% у университетов и научных организаций:

```
SELECT owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (), 0) AS pct
FROM catalogs
WHERE catalog_type = 'Scientific data repository'
GROUP BY 1
ORDER BY n DESC;
```

Страны — США 2018, Германия 764, Великобритания 503; международные записаны как `World`:

```
SELECT country_id, count(*) AS n
FROM (
  SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id
  FROM catalogs
  WHERE catalog_type = 'Scientific data repository'
)
GROUP BY 1
ORDER BY n DESC
LIMIT 20;
```

Платформы с долей API и недействующих — своя разработка 2836 и 5% с API, IPT 360 и 12% недействующих:

```
SELECT software.id AS sid, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       round(100.0 * count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) / count(*),
1) AS pct_dead
FROM catalogs
WHERE catalog_type = 'Scientific data repository'
GROUP BY 1
ORDER BY n DESC
LIMIT 30;
```

Главные платформы по странам — так видно WEKO3 в Японии, DABAR в Хорватии, OPUS в Германии:

```
SELECT country_id, sid, n
FROM (
  SELECT country_id, sid, count(*) AS n,
         row_number() OVER (PARTITION BY country_id ORDER BY count(*) DESC) AS rn
  FROM (
    SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id,
           software.id AS sid
    FROM catalogs
    WHERE catalog_type = 'Scientific data repository'
  )
  GROUP BY 1, 2
)
WHERE rn <= 4 AND country_id IN ('US', 'DE', 'GB', 'JP', 'CN', 'HR', 'FR', 'RU', 'World')
ORDER BY country_id, n DESC;
```

Интерфейсы — OAI-PMH у 2068, карта сайта у 2029, DCAT около 200:

```
SELECT e.type, count(DISTINCT id) AS n
FROM (
  SELECT id, unnest(endpoints) AS e
  FROM catalogs
  WHERE catalog_type = 'Scientific data repository'
)
GROUP BY 1
ORDER BY n DESC
LIMIT 25;
```

Темы — окружающая среда около 2,3 тысячи, живые организмы около 1,6 тысячи. Один репозиторий даёт несколько строк, темы взяты из двух справочников:

```
SELECT t.type, t.name, count(DISTINCT id) AS n
FROM (
  SELECT id, unnest(topics) AS t
  FROM catalogs
  WHERE catalog_type = 'Scientific data repository'
)
GROUP BY 1, 2
ORDER BY n DESC
LIMIT 20;
```

Каталоги показателей

Глава «Каталоги показателей». Число, доля самописных, доля API и недействующие — 3405, 44%, 37% и 29:

```
SELECT count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE software.id = 'custom') / count(*), 0) AS
pct_custom,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead
FROM catalogs
WHERE catalog_type = 'Indicators catalog';
```

Владельцы — 46% у центральных органов власти; доля самописных и API внутри каждой группы:

```
SELECT owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (), 0) AS pct,
       round(100.0 * count(*) FILTER (WHERE software.id = 'custom') / count(*), 0) AS
pct_custom,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api
FROM catalogs
WHERE catalog_type = 'Indicators catalog'
GROUP BY 1
ORDER BY n DESC;
```

Платформы с долей API — Swing 214, Knoema 125, PxWeb 110, DHIS2 74. У WordPress и Drupal поле `api` отмечает API самой CMS, а не рядов:

```
SELECT software.id AS sid, count(*) AS n,
```

```
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       round(100.0 * count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) / count(*),
1) AS pct_dead
FROM catalogs
WHERE catalog_type = 'Indicators catalog'
GROUP BY 1
ORDER BY n DESC
LIMIT 30;
```

Страны у каждой платформы — так видно, что Swing — это Нидерланды и Бельгия, HCI — США, Géoclip — Франция, Knoema — Африка:

```
WITH t AS (
  SELECT software.id AS sid,
         unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id
  FROM catalogs
  WHERE catalog_type = 'Indicators catalog'
    AND software.id IN ('swing', 'knoema', 'pxweb', 'dhis2', 'opensdg', 'hci', 'geoclip',
'statsuite', 'sdmxri')
)
SELECT sid, string_agg(country_id || ':' || n, ', ' ORDER BY n DESC) AS countries
FROM (SELECT sid, country_id, count(*) AS n FROM t GROUP BY 1, 2)
GROUP BY 1;
```

Каталоги с адресами SDMX — 170; адреса DCAT встречаются меньше чем у десятка:

```
SELECT count(*) FILTER (WHERE list_has_any(list_transform(endpoints, lambda e: e.type),
         ['sdmx:datastructure', 'sdmx:dataflows', 'sdmx-rest', 'sdmx:codelist',
'sdmx:conceptscheme'])) AS with_sdmx
FROM catalogs
WHERE catalog_type = 'Indicators catalog';

SELECT e.type, count(*) AS n
FROM (SELECT unnest(endpoints) AS e FROM catalogs WHERE catalog_type = 'Indicators catalog')
t
WHERE lower(e.type) LIKE '%dcat%'
GROUP BY 1;
```

Поисковики и агрегаторы

Глава «Поисковики и агрегаторы». Владельцы — 68% у университетов и научных организаций. В главе «Сообщества и общественные организации» — это `Community` и `Civil society`, «Прочие» — `Regional government` и `Other`:

```
SELECT owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (), 0) AS pct
FROM catalogs
WHERE catalog_type = 'Data search engine'
GROUP BY 1
ORDER BY n DESC;
```

Территории — Евросоюз целиком 37, Италия 23, весь мир 21:

```
SELECT country_id, count(*) AS n
FROM (
  SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id
  FROM catalogs
```

```
  WHERE catalog_type = 'Data search engine'
)
GROUP BY 1
ORDER BY n DESC
LIMIT 20;
```

Платформы с долей API и недействующих — OpenAIRE 50, своя разработка 35 и 20% с API, VIVO 28, Idra 6 и все недействующие:

```
SELECT software.id AS sid, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead
FROM catalogs
WHERE catalog_type = 'Data search engine'
GROUP BY 1
ORDER BY n DESC;
```

Полный список — так видны 17 итальянских UNIFIND на VIVO и демонстрационные стенды Idra:

```
SELECT software.id AS sid, owner.type AS owner_type, status, name, link
FROM catalogs
WHERE catalog_type = 'Data search engine'
ORDER BY sid, name;
```

Крупные поисковики, записанные другими типами, — Dateno, NIAID Discovery Portal, Research Data Australia, GESIS Search, DataONE, B2FIND, Earthdata Search, OmicsDI:

```
SELECT catalog_type, software.id AS sid, status, name, link
FROM catalogs
WHERE regexp_matches(lower(link),

'(dateno\.io|//data\.niaid|researchdata\.edu\.au|search\.gesis|//www\.dataone\.org|b2find|sea
rch\.earthdata|omicsdi)')
ORDER BY catalog_type, name;
```

Каталоги метаданных

Глава «Каталоги метаданных и реестры». Число, доля самописных, доля API, недействующие и с ограниченным доступом — 172, 38%, 67%, 0 и 18:

```
SELECT count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE software.id = 'custom') / count(*), 0) AS
pct_custom,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) AS restricted
FROM catalogs
WHERE catalog_type = 'Metadata catalog';
```

Владельцы — 31% у университетов и научных организаций, 30% у международных организаций. В главе `Community` и `Other` сложены в «Сообщества и прочие»:

```
SELECT owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (), 0) AS pct
```

```
FROM catalogs
WHERE catalog_type = 'Metadata catalog'
GROUP BY 1
ORDER BY n DESC;
```

Платформы с долей API — своя разработка 65 и 17%, FAIR Data Point 47, Metadata Browser (`mwmb`) 16, Skosmos 10:

```
SELECT software.id AS sid, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api
FROM catalogs
WHERE catalog_type = 'Metadata catalog'
GROUP BY 1
ORDER BY n DESC;
```

Платформы по странам — так видно, что Нидерланды — это точки FAIR Data Point, Великобритания — Metadata Browser, Австралия — Aristotle:

```
SELECT country_id, sid, count(*) AS n
FROM (
  SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id,
         software.id AS sid
  FROM catalogs
  WHERE catalog_type = 'Metadata catalog'
)
GROUP BY 1, 2
ORDER BY country_id, n DESC;
```

Интерфейсы — DCAT в Turtle у 47 точек FAIR Data Point, REST API Skosmos у 9 установок, адреса SDMX у реестров на Fusion Metadata Registry:

```
SELECT software.id AS sid, e.type, count(DISTINCT id) AS n
FROM (
  SELECT id, software, unnest(endpoints) AS e
  FROM catalogs
  WHERE catalog_type = 'Metadata catalog'
)
GROUP BY 1, 2
ORDER BY 1, n DESC;
```

ML-каталоги

Глава «ML-каталоги». Всего 100, недействующих 2, с ограниченным доступом 5, самописных 78%, с API 23%:

```
SELECT count(*) AS n,
       count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) AS restricted,
       round(100.0 * count(*) FILTER (WHERE software.id = 'custom') / count(*), 0) AS
pct_custom,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api
FROM catalogs
WHERE catalog_type = 'Machine learning catalog';
```

Владельцы — 37 у университетов и научных организаций, 31 у компаний:

```
SELECT owner.type, count(*) AS n
FROM catalogs
WHERE catalog_type = 'Machine learning catalog'
GROUP BY 1
ORDER BY n DESC;
```

Страны — международные 34, США 19, Китай 12, Россия 6:

```
SELECT country_id, count(*) AS n
FROM (
  SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id
  FROM catalogs
  WHERE catalog_type = 'Machine learning catalog'
)
GROUP BY 1
ORDER BY n DESC
LIMIT 15;
```

Платформы и доля API — своя разработка 78 и 13 с API, Jekyll 9:

```
SELECT software.id AS sid, count(*) AS n,
       count(*) FILTER (WHERE api) AS with_api
FROM catalogs
WHERE catalog_type = 'Machine learning catalog'
GROUP BY 1
ORDER BY n DESC;
```

Полный список записей. По нему в главе записи вручную разложены на восемь видов — хабы, площадки соревнований, коллекции, сайты бенчмарков и другие; такого поля в реестре нет:

```
SELECT name, link, software.id AS sid, owner.type AS owner_type, api, status
FROM catalogs
WHERE catalog_type = 'Machine learning catalog'
ORDER BY sid, name;
```

Каталоги API, маркетплейсы и списки наборов

Глава «Каталоги API, маркетплейсы и списки наборов». Число, доля своих разработок, API и ограниченного доступа — 115, 74%, 51% и 14% у каталогов API; 56, 98%, 21% и 91% у маркетплейсов; 107, 89%, 14% и 1% у списков:

```
SELECT catalog_type, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE software.id = 'custom') / count(*), 0) AS
pct_custom,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       round(100.0 * count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) /
count(*), 0) AS pct_restricted
FROM catalogs
WHERE catalog_type IN ('API Catalog', 'Data marketplace', 'Datasets list')
GROUP BY 1;
```

Владельцы — 45% каталогов API у центральных органов власти, 70% маркетплейсов у компаний:

```
SELECT catalog_type, owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (PARTITION BY catalog_type), 0) AS pct
FROM catalogs
WHERE catalog_type IN ('API Catalog', 'Data marketplace', 'Datasets list')
GROUP BY 1, 2
ORDER BY 1, n DESC;
```

Страны — у маркетплейсов Южная Корея 16, США 12, Китай 6:

```
SELECT catalog_type, country_id, count(*) AS n
FROM (
  SELECT catalog_type,
         unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id
  FROM catalogs
  WHERE catalog_type IN ('API Catalog', 'Data marketplace', 'Datasets list')
)
GROUP BY 1, 2
QUALIFY row_number() OVER (PARTITION BY catalog_type ORDER BY count(*) DESC) <= 10
ORDER BY 1, n DESC;
```

Платформы — у каталогов API Azure API Management 22, WSO2 API Manager 2:

```
SELECT catalog_type, software.id AS sid, count(*) AS n,
       count(*) FILTER (WHERE api) AS with_api
FROM catalogs
WHERE catalog_type IN ('API Catalog', 'Data marketplace', 'Datasets list')
GROUP BY 1, 2
ORDER BY 1, n DESC;
```

Транспортные порталы разработчиков, около 35 из 115, отобраны по названию и адресу; список стоит просмотреть глазами, потому что шаблон ловит и лишнее (Metropolia, gov.br/conecta):

```
SELECT name, link
FROM catalogs
WHERE catalog_type = 'API Catalog'
  AND regexp_matches(lower(name || ' ' || link),

'transit|transport|metro|\bbus\b|rail|traffic|trafik|tmb|mta|bart|cta|marta|wmata|trimet|tran
slink|entur|waltti|föli|foli|metlink|511|digitraffic|tdx|koroad|apiportal.ns|sptrans|oc
transpo|skåne|cumtd|cota|deutschebahn');
```

Списки наборов на GitHub и GitLab — 16, история торгов криптобирж — 13:

```
SELECT
  count(*) FILTER (WHERE regexp_matches(lower(name || ' ' || link),
'github.io|github.com|gitlab')) AS on_git,
  count(*) FILTER (WHERE regexp_matches(lower(name || ' ' || link),

'binance|bybit|okx|kucoin|mexc|bitget|kraken|gatedata|bitmex|crypto|polymarket|paradigm')) AS
crypto
FROM catalogs
WHERE catalog_type = 'Datasets list';
```

Карта сайта у каталогов API — 26 из 115:

```
SELECT catalog_type, e.type, count(DISTINCT id) AS n
FROM (
  SELECT id, catalog_type, unnest(endpoints) AS e
  FROM catalogs
  WHERE catalog_type IN ('API Catalog', 'Data marketplace', 'Datasets list')
)
GROUP BY 1, 2
ORDER BY 1, n DESC;
```

Карта возможностей

Глава «Карта возможностей», запросы 15 октября 2026 года. Интерфейсы — по списку `endpoints`, DOI — по признаку `has_doi`; научные репозитории посчитаны вместе с универсальными. Разметку schema.org реестр почти не записывает, поэтому её доля не считается.

Доли по типам — DCAT у 26% порталов открытых данных, OAI-PMH у 25% и DOI у 17% научных репозиториев, CSW или OGC API Records у 20% геопорталов, SDMX у 5,2% каталогов показателей:

```
WITH t AS (
  SELECT CASE WHEN catalog_type IN ('Scientific data repository', 'General research
repository')
              THEN 'Science' ELSE catalog_type END AS ct,
         properties.has_doi AS doi,
         list_transform(coalesce(endpoints, []), lambda e: lower(e.type)) AS et
  FROM catalogs
)
SELECT ct, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE 'dcat%' OR x
LIKE '%:dcatus11')) > 0) / count(*), 1) AS pct_dcat,
       round(100.0 * count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE 'oai%')) >
0) / count(*), 1) AS pct_oai,
       round(100.0 * count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE 'csw%' OR x
LIKE 'ogcrecords%')) > 0) / count(*), 1) AS pct_csw_records,
       round(100.0 * count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE '%sdmx%')) >
0) / count(*), 1) AS pct_sdmx,
       round(100.0 * count(*) FILTER (WHERE doi) / count(*), 1) AS pct_doi
FROM t
WHERE ct IN ('Open data portal', 'Science', 'Geoportal', 'Indicators catalog')
GROUP BY 1
ORDER BY n DESC;
```

DCAT у порталов открытых данных по платформам — Socrata 239 из 245, OpenDataSoft 348 из 454, DKAN 103 из 161, CKAN 633 из 1906, своя разработка 24 из 1144:

```
SELECT software.id AS sid, count(*) AS n,
       count(*) FILTER (WHERE len(list_filter(
         list_transform(coalesce(endpoints, []), lambda e: lower(e.type)),
         lambda x: x LIKE 'dcat%' OR x LIKE '%:dcatus11')) > 0) AS with_dcat
FROM catalogs
WHERE catalog_type = 'Open data portal'
GROUP BY 1
ORDER BY n DESC
LIMIT 10;
```

DOI и OAI-PMH у научных репозиториев по платформам — Figshare 266 DOI из 288, Dataverse 172 DOI и 177 OAI-PMH из 257, DSpace 218 DOI и 665 OAI-PMH из 829, своя разработка 158 DOI и 19 OAI-PMH из 2840:

```
SELECT software.id AS sid, count(*) AS n,
       count(*) FILTER (WHERE properties.has_doi) AS with_doi,
       count(*) FILTER (WHERE len(list_filter(
         list_transform(coalesce(endpoints, []), lambda e: lower(e.type)),
         lambda x: x LIKE 'oai%')) > 0) AS with_oai
FROM catalogs
WHERE catalog_type IN ('Scientific data repository', 'General research repository')
GROUP BY 1
ORDER BY n DESC
LIMIT 10;
```

Каталожный сервис у геопорталов по платформам — ArcGIS Hub 3257 из 3464, GeoNetwork 686 из 762, GeoNode 628 из 770, ArcGIS Server 1 из 4674, e-mapa 0 из 1927, своя разработка 4 из 1043:

```
SELECT software.id AS sid, count(*) AS n,
       count(*) FILTER (WHERE len(list_filter(
         list_transform(coalesce(endpoints, []), lambda e: lower(e.type)),
         lambda x: x LIKE 'csw%' OR x LIKE 'ogcrecords%')) > 0) AS with_catalog_service
FROM catalogs
WHERE catalog_type = 'Geoportal'
GROUP BY 1
ORDER BY n DESC
LIMIT 10;
```

SDMX у каталогов показателей по платформам — из 177 каталогов 113 на Knoema:

```
SELECT software.id AS sid, count(*) AS n
FROM catalogs
WHERE catalog_type = 'Indicators catalog'
  AND len(list_filter(
        list_transform(coalesce(endpoints, []), lambda e: lower(e.type)),
        lambda x: x LIKE '%sdmx%')) > 0
GROUP BY 1
ORDER BY n DESC;
```

Экспорт метаданных

Глава «Экспорт метаданных», запросы 15 октября 2026 года. Хотя бы один стандартный способ отдать описания — DCAT, OAI-PMH, CSW, OGC API Records, SDMX или STAC — у 9 330 каталогов из 43 614. Выгрузка DCAT есть у 5 233, сразу оба профиля, DCAT-US и DCAT-AP, — у 3 152. Разметку schema.org реестр почти не записывает, таких адресов 38:

```
WITH f AS (
  SELECT list_transform(coalesce(endpoints, []), lambda e: lower(e.type)) AS et
  FROM catalogs
)
SELECT count(*) AS n,
       count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE 'dcat%' OR x LIKE
'%dcatus11'
                        OR x = 'dcat-us')) > 0) AS dcat,
```

```
       count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE 'dcatus11' OR x LIKE
'%:dcatus11'
                        OR x = 'dcat-us')) > 0
                        AND len(list_filter(et, lambda x: x LIKE 'dcatap%')) > 0) AS
both_profiles,
       count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE 'dcat%' OR x LIKE
'%dcatus11'
                        OR x = 'dcat-us' OR x LIKE 'oai%' OR x LIKE 'csw%' OR x LIKE
'ogcrecords%'
                        OR x LIKE '%sdmx%' OR x LIKE 'stac%')) > 0) AS any_standard,
       count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE 'schemaorg%')) > 0) AS
schemaorg
FROM f;
```

DCAT у порталов открытых данных по платформам — Socrata 239 из 245, OpenDataSoft 348 из 454, DKAN 103 из 161, CKAN 633 из 1906, своя разработка 24 из 1144. Запрос приведён в разделе «Карта возможностей».

API и массовая выгрузка

Глава «API и массовая выгрузка», запросы 15 октября 2026 года. Признак «есть программный вход» в реестре совпадает с непустым списком `endpoints`: 28 471 из 43 614. Стандартный способ забрать все описания — DCAT, OAI-PMH, CSW или OGC API Records — у 8 894 каталогов. Доли по типам — 70% геопорталов, 69% порталов открытых данных, 63% научных репозиториев, 37% каталогов показателей, 23% ML-каталогов:

```
WITH t AS (
  SELECT catalog_type,
         list_transform(coalesce(endpoints, []), lambda e: lower(e.type)) AS et
  FROM catalogs
), f AS (
  SELECT catalog_type,
         len(et) > 0 AS any_ep,
         len(list_filter(et, lambda x: x LIKE 'dcat%' OR x LIKE '%dcatus11' OR x LIKE 'oai%'
                                    OR x LIKE 'csw%' OR x LIKE 'ogcrecords%')) > 0 AS
all_desc,
         len(list_filter(et, lambda x: x LIKE '%sitemap%')) > 0 AS sitemap
  FROM t
)
SELECT coalesce(catalog_type, 'всего') AS catalog_type, count(*) AS n,
       count(*) FILTER (WHERE any_ep) AS any_ep,
       round(100.0 * count(*) FILTER (WHERE any_ep) / count(*), 0) AS pct_any,
       count(*) FILTER (WHERE all_desc) AS all_desc,
       round(100.0 * count(*) FILTER (WHERE all_desc) / count(*), 0) AS pct_all_desc,
       round(100.0 * count(*) FILTER (WHERE sitemap) / count(*), 0) AS pct_sitemap
FROM f
GROUP BY ROLLUP (catalog_type)
ORDER BY n DESC;
```

Своя разработка против готовых платформ — 697 из 7 034 с программным входом и 58 со стандартной выгрузкой описаний у своих разработок, 76% с программным входом у остальных:

```
WITH t AS (
  SELECT software.id = 'custom' AS is_custom,
         list_transform(coalesce(endpoints, []), lambda e: lower(e.type)) AS et
  FROM catalogs
```

```
)
SELECT is_custom, count(*) AS n,
       count(*) FILTER (WHERE len(et) > 0) AS any_ep,
       round(100.0 * count(*) FILTER (WHERE len(et) > 0) / count(*), 0) AS pct_any,
       count(*) FILTER (WHERE len(list_filter(et, lambda x: x LIKE 'dcat%' OR x LIKE
'%dcatus11'
                        OR x LIKE 'oai%' OR x LIKE 'csw%' OR x LIKE 'ogcrecords%')) > 0) AS
all_desc
FROM t
GROUP BY 1;
```

Ленты RSS и Atom реестр записывает в основном у геопорталов и научных репозиториев, у порталов на CKAN их почти нет, хотя лента в ядре CKAN. Поэтому доля лент по реестру в главе не приводится.

Идентификаторы и DOI

Глава «Постоянные идентификаторы, DOI и версии», запросы 15 октября 2026 года. Пометка DOI — у 2008 каталогов из 43 614: 1396 научных репозиториев и 8 универсальных, 584 геопортала, остальные единичны:

```
SELECT catalog_type,
       count(*) FILTER (WHERE properties.has_doi) AS with_doi,
       count(*) AS n
FROM catalogs
GROUP BY 1
HAVING with_doi > 0
ORDER BY with_doi DESC;
```

Пометку в основном наследуют от карточки платформы (`software.pid_support.has_doi`), поэтому её надо читать как «платформа умеет». У GeoNetwork платформа отмечена «Yes», и пометку получили 555 установок из 764; у своих разработок платформа отмечена «No», и 175 пометок поставлены вручную:

```
SELECT c.software.id AS sid,
       s.pid_support.has_doi AS platform_flag,
       count(*) FILTER (WHERE c.properties.has_doi) AS with_doi,
       count(*) AS n
FROM catalogs c
LEFT JOIN software s ON s.id = c.software.id
GROUP BY 1, 2
HAVING with_doi > 0
ORDER BY with_doi DESC
LIMIT 10;
```

Выборка для проверки Handle — 120 из 782 адресов OAI-PMH у действующих установок DSpace, в воспроизводимом случайном порядке. Для каждой у первой записи из `ListIdentifiers` берут Handle и проверяют его через `https://hdl.handle.net/api/handles/…`:

```
SELECT e.url
FROM (SELECT unnest(endpoints) AS e
      FROM catalogs
      WHERE software.id = 'dspace' AND status = 'active')
WHERE lower(e.type) LIKE 'oai%'
```

```
ORDER BY hash(e.url)
LIMIT 120;
```

Доступ, роли и процессы

Глава «Доступ, роли и процессы публикации», запросы 15 октября 2026 года. Ограниченный доступ хотя бы к части данных — у 1131 каталога из 43 614: у 537 только ограниченный, у 594 вместе с открытым:

```
SELECT count(*) AS total,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) AS restricted,
       count(*) FILTER (WHERE access_mode = ['restricted']) AS only_restricted,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')
                          AND list_contains(access_mode, 'open')) AS mixed
FROM catalogs;
```

Доля по платформам — DHIS2 97%, NADA 72%, Pure 36% против CKAN 0,6% и Dataverse 1,9%:

```
SELECT software.id AS platform,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) AS restricted,
       count(*) AS n,
       round(100.0 * restricted / n, 1) AS pct
FROM catalogs
WHERE software.id IN ('nada', 'dhis2', 'gen3', 'redivis', 'pure', 'ckan', 'dataverse',
                      'dspace', 'inveniordm', 'geonetwork', 'opendatasoft', 'socrata',
'arcgishub')
GROUP BY 1
ORDER BY pct DESC;
```

Список для замера регистрации на порталах CKAN — 1839 действующих каталогов. Для каждого берут адрес API из точки входа `ckan`, а если её нет — адрес сайта с `/api/3`:

```
SELECT id, link, catalog_type, owner.type AS otype,
       coverage[1].location.country.id AS country,
       [x.url for x in endpoints if x.type = 'ckan'][1] AS api
FROM catalogs
WHERE software.id = 'ckan' AND status = 'active'
ORDER BY id;
```

Каталоги микроданных

Глава «Каталоги микроданных». Всего 326, неработающих 59, с ограниченным доступом 159, самописных 13%, с API 70%:

```
SELECT count(*) AS n,
       count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) AS restricted,
       round(100.0 * count(*) FILTER (WHERE software.id = 'custom') / count(*), 0) AS
pct_custom,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api
FROM catalogs
WHERE catalog_type = 'Microdata catalog';
```

Владельцы и доля NADA у каждого — 55% у центральных органов власти, из них 59% на NADA; у международных организаций 90% на NADA:

```
SELECT owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (), 0) AS pct,
       round(100.0 * count(*) FILTER (WHERE software.id = 'nada') / count(*), 0) AS pct_nada
FROM catalogs
WHERE catalog_type = 'Microdata catalog'
GROUP BY 1
ORDER BY n DESC;
```

Платформы с долей API и числом неработающих — NADA 158 и 47 неработающих, своя разработка 43 и 12% с API:

```
SELECT software.id AS sid, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead,
       count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) AS restricted
FROM catalogs
WHERE catalog_type = 'Microdata catalog'
GROUP BY 1
ORDER BY n DESC;
```

Страны — 139 стран и 27 общемировых каталогов; США 18, Канада 12, Германия 11:

```
SELECT country_id, count(*) AS n
FROM (
  SELECT unnest(list_transform(coverage, lambda x: x.location.country.id)) AS country_id
  FROM catalogs
  WHERE catalog_type = 'Microdata catalog'
)
GROUP BY 1
ORDER BY n DESC;
```

Неработающие каталоги без HTTPS и на IP-адресах — 38 из 58 без HTTPS, 7 из 10 на IP-адресах:

```
SELECT count(*) FILTER (WHERE link LIKE 'http://%') AS http,
       count(*) FILTER (WHERE link LIKE 'http://%' AND status = 'inactive') AS http_dead,
       count(*) FILTER (WHERE regexp_matches(link, '^https?://[0-9.]+')) AS ip,
       count(*) FILTER (WHERE regexp_matches(link, '^https?://[0-9.]+') AND status =
'inactive') AS ip_dead
FROM catalogs
WHERE catalog_type = 'Microdata catalog';
```

Страны с неработающей NADA и без другого действующего каталога микроданных — 18 из 38:

```
WITH c AS (
  SELECT status, software.id AS sid,
         unnest(list_transform(coverage, lambda x: x.location.country.id)) AS cc
  FROM catalogs
  WHERE catalog_type = 'Microdata catalog'
),
dead AS (SELECT DISTINCT cc FROM c WHERE status = 'inactive' AND sid = 'nada' AND length(cc)
= 2)
SELECT count(*) AS countries,
```

```
       count(*) FILTER (WHERE NOT EXISTS (
         SELECT 1 FROM c WHERE c.cc = dead.cc AND c.status = 'active')) AS without_active
FROM dead;
```

Сравнение типов и гибриды

Глава «Сравнение типов и гибриды». Сводная таблица — число, доли своих разработок, API, неработающих и ограниченного доступа по типам:

```
SELECT catalog_type, count(*) AS n,
       round(100.0 * count(*) FILTER (WHERE software.id = 'custom') / count(*), 0) AS
pct_custom,
       round(100.0 * count(*) FILTER (WHERE api) / count(*), 0) AS pct_api,
       round(100.0 * count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) / count(*),
1) AS pct_dead,
       round(100.0 * count(*) FILTER (WHERE list_contains(access_mode, 'restricted')) /
count(*), 0) AS pct_restricted
FROM catalogs
GROUP BY 1
ORDER BY n DESC;
```

Главный владелец и четыре главные платформы каждого типа, без своих разработок:

```
SELECT catalog_type, owner.type, count(*) AS n,
       round(100.0 * count(*) / sum(count(*)) OVER (PARTITION BY catalog_type), 0) AS pct
FROM catalogs
GROUP BY 1, 2
QUALIFY row_number() OVER (PARTITION BY catalog_type ORDER BY count(*) DESC) = 1;

SELECT catalog_type, string_agg(sid || ' ' || n, ', ' ORDER BY n DESC) AS top
FROM (
  SELECT catalog_type, software.id AS sid, count(*) AS n
  FROM catalogs
  WHERE software.id <> 'custom'
  GROUP BY 1, 2
  QUALIFY row_number() OVER (PARTITION BY catalog_type ORDER BY count(*) DESC) <= 4
)
GROUP BY 1;
```

В скольких типах встречается платформа — WordPress в девяти, CKAN в пяти, GeoNode, Dataverse и PxWeb в одном:

```
SELECT software.id AS sid, count(DISTINCT catalog_type) AS n_types, count(*) AS n,
       string_agg(DISTINCT catalog_type, ', ') AS types
FROM catalogs
WHERE catalog_type NOT IN ('Other', 'General research repository')
  AND software.id <> 'custom'
GROUP BY 1
ORDER BY n_types DESC, n DESC
LIMIT 30;
```

Владельцы с каталогами нескольких типов — 2 203 из 32 595, из них 254 с тремя и более, за ними 7 358 каталогов; одной программой обходятся 118. Владелец здесь — название организации в одной стране, поэтому подсчёт приблизительный:

```
CREATE TEMP TABLE o AS
```

```
SELECT lower(trim(owner.name)) AS oname, owner.location.country.id AS cc,
       owner.type AS otype, catalog_type, software.id AS sid
FROM catalogs
WHERE coalesce(trim(owner.name), '') <> ''
  AND catalog_type NOT IN ('Other', 'General research repository');

CREATE TEMP TABLE g AS
SELECT oname, cc, any_value(otype) AS otype,
       count(DISTINCT catalog_type) AS nt, count(*) AS n, count(DISTINCT sid) AS ns,
       string_agg(DISTINCT catalog_type, ' | ' ORDER BY catalog_type) AS type_list
FROM o
GROUP BY 1, 2;

SELECT count(*) AS owners,
       count(*) FILTER (WHERE nt > 1) AS multi,
       count(*) FILTER (WHERE nt > 2) AS multi3,
       sum(n) FILTER (WHERE nt > 1) AS catalogs,
       count(*) FILTER (WHERE nt > 1 AND ns = 1) AS one_software
FROM g;
```

Доля таких владельцев по виду владельца и самые частые пары типов — геопортал и портал открытых данных у 953:

```
SELECT otype, count(*) AS owners,
       round(100.0 * count(*) FILTER (WHERE nt > 1) / count(*), 0) AS pct_multi
FROM g
GROUP BY 1
ORDER BY owners DESC;

SELECT type_list, count(*) AS owners
FROM g
WHERE nt = 2
GROUP BY 1
ORDER BY owners DESC
LIMIT 15;
```

Пара «геопортал и портал открытых данных»: 1 120 владельцев, сочетания программ (ArcGIS Server и CKAN — 101) и одна программа на обеих сторонах — у 44 из 953, из них 39 своих разработок:

```
CREATE TEMP TABLE pair AS
SELECT oname, cc, any_value(otype) AS otype
FROM o
GROUP BY 1, 2
HAVING bool_or(catalog_type = 'Geoportal') AND bool_or(catalog_type = 'Open data portal');

SELECT g.sid AS geo, d.sid AS odp, count(DISTINCT oname || cc) AS owners
FROM (SELECT DISTINCT oname, cc, sid FROM o JOIN pair USING (oname, cc) WHERE catalog_type =
'Geoportal') g
JOIN (SELECT DISTINCT oname, cc, sid FROM o JOIN pair USING (oname, cc) WHERE catalog_type =
'Open data portal') d
  USING (oname, cc)
GROUP BY 1, 2
ORDER BY owners DESC
LIMIT 12;

SELECT shared_sid, count(*) AS owners
FROM (
  SELECT oname, cc,
         unnest(list_intersect(
```

```
           list(DISTINCT sid) FILTER (WHERE catalog_type = 'Geoportal'),
           list(DISTINCT sid) FILTER (WHERE catalog_type = 'Open data portal'))) AS
shared_sid
  FROM o JOIN g USING (oname, cc)
  WHERE g.nt = 2 AND g.type_list = 'Geoportal | Open data portal'
  GROUP BY 1, 2
)
GROUP BY 1
ORDER BY owners DESC;
```

Каталоги одной организации — так видны составы Тихоокеанского сообщества, Всемирного банка, ФАО и INEGI:

```
SELECT owner.name, catalog_type, software.id AS sid, status, link
FROM catalogs
WHERE owner.name IN ('Pacific Community (SPC)', 'World Bank',
                     'Food and Agriculture Organization of the United Nations',
                     'Instituto Nacional de Estadística y Geografía (INEGI)')
ORDER BY 1, 2, 3;
```

Когда портал не нужен

Глава «Когда портал не нужен», запросы 15 октября 2026 года к той же выгрузке. Списки наборов — 107, программный интерфейс у 15, своя разработка у 95, недействующие и снятые — 7:

```
SELECT
  count(*) AS n,
  count(*) FILTER (WHERE api) AS with_api,
  count(*) FILTER (WHERE software.id = 'custom') AS custom,
  count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead
FROM catalogs
WHERE catalog_type = 'Datasets list';
```

Карта сайта — у 12. Остальные движки страниц — Drupal, WordPress, TYPO3, Jekyll, SharePoint, Liferay:

```
SELECT count(DISTINCT id) AS with_sitemap
FROM (
  SELECT id, unnest(endpoints) AS e
  FROM catalogs
  WHERE catalog_type = 'Datasets list'
)
WHERE e.type = 'sitemap';

SELECT software.id AS sid, count(*) AS n
FROM catalogs
WHERE catalog_type = 'Datasets list'
GROUP BY 1
ORDER BY n DESC;
```

Передовые каталоги

Глава «Что уже умеют передовые каталоги», запросы 15 октября 2026 года к той же выгрузке. ERDDAP — 175 записей, недействующая одна. Datasette — 45, недействующая одна. Redivis — 13, DataLad — 10, DANDI — 3. Datadex — 3, программный интерфейс не отмечен ни у одной:

```
SELECT software.id AS sid, count(*) AS n,
       count(*) FILTER (WHERE status IN ('inactive', 'deprecated')) AS dead,
       count(*) FILTER (WHERE api) AS with_api
FROM catalogs
WHERE software.id IN ('erddap', 'datasette', 'redivis', 'datalad', 'dandi', 'datadex')
GROUP BY 1
ORDER BY n DESC;
```

Записи с openEO в адресе или типе точки доступа — 12, из них у Copernicus Data Space Ecosystem — 3. Каталоги с интерфейсом GA4GH DRS — 27:

```
SELECT count(DISTINCT id) AS with_openeo,
       count(DISTINCT id) FILTER (WHERE e.url ILIKE '%dataspace.copernicus.eu%') AS cdse
FROM (SELECT id, unnest(endpoints) AS e FROM catalogs)
WHERE e.url ILIKE '%openeo%' OR e.type ILIKE 'openeo%';

SELECT count(DISTINCT id) AS with_drs
FROM (SELECT id, unnest(endpoints) AS e FROM catalogs)
WHERE e.type = 'ga4gh:drs';
```

В число с openEO попадают и два сайта сообщества с одной картой сайта: адрес содержит слово, а сервера openEO за ним нет. Записи для сверки с главой — Source Cooperative, Data Commons, Our World in Data, DoltHub:

```
SELECT id, catalog_type, software.id AS sid, api, link
FROM catalogs
WHERE link ILIKE '%source.coop%' OR link ILIKE '%datacommons.org%'
   OR link ILIKE '%ourworldindata.org%' OR link ILIKE '%dolthub.com%';
```

В тот день у DoltHub и Our World in Data из точек доступа записана только карта сайта, хотя у обоих есть API. У Data Commons точек доступа нет вовсе и программный интерфейс не отмечен, хотя у него есть REST, SDMX и сервер MCP. У основной записи Source Cooperative — один интерфейс STAC, у записи `beta.source.coop` точек доступа нет.

Дальше

- Примеры запросов в документации реестра
- Модель данных

Сравнительные таблицы платформ

Как читать клетки. В клетке одна из четырёх пометок. Текст в скобках — не пятая пометка, а имя модуля или оговорка.

- **ядро** — возможность есть в самой платформе;
- **расширение** — нужен отдельный модуль; кто его ведёт и с какой версией ядра он работает, написано в карточке;
- **нет** — платформа этого не делает;
- **не проверено** — по документации в октябре 2026 года клетку нельзя было закрыть. Перед выбором платформы проверьте её сами.

Облачные сервисы не принимают чужие расширения, поэтому у них нет клетки «расширение». Строка «свои расширения» отвечает на другой вопрос: можно ли вообще добавить модуль. Подробности — в карточках. Те же четыре пометки используются в главах о возможностях. Слова «доработка» и «невозможно» встречаются только в сравнении короткого списка: они значат «придётся писать самим» и «на этой платформе путь закрыт».

Открытые данные и лёгкие каталоги

В реестре Huwise записана старым именем OpenDataSoft. Здесь — новое имя, старое в скобках. Сбор у Huwise и Socrata сверен с документацией поставщиков 14 октября 2026 года. Остальные клетки — по главам о возможностях.

Возможность	CKAN	DKAN	Huwise (OpenDataSoft)	Socrata	ArcGIS Hub	JKAN
API поиска	ядро	ядро	ядро	ядро	ядро	нет
Выгрузка всех описаний	расширение (ckanext-dcat)	ядро	ядро	ядро	ядро	ядро (файл data.json)
Лента и выборка изменений	ядро (лента Atom, фильтр по дате изменения)	не проверено	ядро (лента RSS, фильтр по дате)	не проверено	ядро (лента RSS)	нет (лента изменений репозитория, если сайт в Git)
DCAT	расширение (ckanext-dcat, по умолчанию профиль DCAT-AP 3)	ядро (файл data.json; схема ядра — DCAT-US 1.1)	ядро (RDF, Turtle и DCAT; национальный профиль бывает не у всех)	ядро (файл data.json по DCAT-US)	ядро (DCAT-US 1.1 и DCAT-AP 2.1.1; DCAT-US 3.0 и DCAT-AP 3.0.0 включаются отдельно)	ядро (файл data.json объявляет DCAT-US 1.1; шаблон заполняет короткий список полей)
schema.org	расширение (плагин structured data в ckanext-dcat)	не проверено	ядро, включает администратор	ядро (JSON-LD на странице набора)	нет	нет (шаблон страницы разметку не ставит)
Файлы у себя	ядро	ядро	ядро	ядро	ядро	ядро (файлы рядом с сайтом)

Возможность	CKAN	DKAN	Huwise (OpenDataSoft)	Socrata	ArcGIS Hub	JKAN
Облачное хранилище файлов	расширение (в 2.12 — ckanext-file-keeper-cloud, раньше — ckanext-s3filestore)	не проверено	ядро (хранит поставщик)	ядро (хранит поставщик)	ядро (хранит поставщик)	нет (файлы рядом с сайтом или по ссылке)
Проверка внешних ссылок	расширение (ckanext-archiver, хранит и копии)	не проверено	не проверено	не проверено	не проверено	нет
API к таблицам	ядро; заливка CSV — расширение XLoader	ядро	ядро	ядро	ядро (сервисы ArcGIS)	нет
Свои поля	расширение (ckanext-scheming)	ядро (замена файла JSON Schema)	ядро (шаблоны полей)	ядро (наборы полей)	нет (один стиль из готовых стандартов на организацию)	нет (правка шаблона)
Словарь полей таблицы	ядро (для таблиц в DataStore)	не проверено	ядро	ядро	ядро (описания полей слоя)	нет
DOI	расширение (ckanext-doi; на CKAN 2.11 и 2.12 не тестировалось)	нет	нет	нет	нет	нет
Цитирование	расширение (ckanext-doi только при DOI; ckanext-citeproc)	нет	нет	нет	нет	нет
Роли	ядро (организации, три роли)	ядро (права Drupal)	ядро сервиса	ядро сервиса (пять ролей на портал, три на набор)	ядро сервиса	нет
Проверка перед публикацией	черновик — ядро; проверка куратором — расширение	ядро (состояния); по умолчанию набор публикуется сразу	нет (есть раздельное право публиковать)	ядро (одобрение автоматическое или ручное)	нет	нет
Единый вход	расширение (SAML, OpenID Connect, LDAP)	не проверено	ядро (SAML, OpenID Connect)	ядро (OpenID Connect; SAML через поддержку)	ядро (SAML, OpenID Connect)	нет
Сбор чужих каталогов	расширение (ckanext-harvest; сбор RDF — ckanext-dcat, геокаталоги — ckanext-spatial)	не проверено	ядро (CKAN, Socrata, ArcGIS, ArcGIS Hub, data.json, data.gouv.fr и другие порталы Huwise; CSW включают по запросу в поддержку)	ядро (Catalog Connector: data.json и публичный ArcGIS Server; обмен между доменами Data & Insights)	не проверено	нет
Свои расширения	да	да, через Drupal	нет	нет	нет	правка шаблона

Наука и геоданные

Возможность	Dataverse	InvenioRDM	DSpace	GeoNetwork	NADA	PxWeb
API	ядро	ядро	ядро	ядро	ядро	ядро

Возможность	Dataverse	InvenioRDM	DSpace	GeoNetwork	NADA	PxWeb
Выгрузка описаний	ядро (OAI-PMH)	ядро (OAI-PMH)	ядро (OAI-PMH)	ядро (CSW и OAI-PMH)	ядро (свой API и DDI Codebook; OAI-PMH нет)	ядро (API таблиц, не каталог файлов)
Выборка изменений	ядро (OAI-PMH по дате)	ядро (OAI-PMH по дате)	ядро (OAI-PMH по дате, ленты OpenSearch)	ядро (OAI-PMH по дате)	не проверено	ядро (PxWebApi 2: таблицы, изменённые за N дней)
DCAT	нет в ядре (сторонний экспортёр DCAT 3)	ядро	нет	ядро	нет	нет
schema.org	ядро (с версии 6.10 в страницу встроена краткая разметка Croissant)	ядро	нет	ядро	ядро (JSON-LD)	нет
Файлы у себя	ядро	ядро	ядро	ядро (чаще ссылка на сервис)	ядро	нет
Облачное хранилище файлов	ядро (S3, Swift, Globus; несколько хранилищ)	ядро (S3 с v13)	ядро (S3; несколько хранилищ)	ядро (модули S3, JCloud и CMIS в поставке)	не проверено	нет
DOI	ядро	ядро	ядро	ядро (нужно включить и подключить DataCite)	только поле для DOI, выпущенного отдельно	нет
Цитирование	ядро	ядро	нет (только метатеги для Google Scholar; перенос из DSpace-CRIS не завершён)	ядро (HTML, текст, RIS, BibTeX)	ядро через DataCite, если есть DOI	нет
Свои поля	ядро	ядро (настройка)	ядро (настройка)	ядро (ISO 19139, ISO 19115-3, Dublin Core); профиль страны — расширение (плагин схемы)	не проверено	нет
Словарь полей таблицы	ядро (переменные загруженных таблиц)	нет	нет	ядро (каталог объектов ISO 19110)	ядро (переменные DDI)	ядро (переменные и значения таблицы)
Ограниченный доступ к файлам	ядро	ядро	ядро	ядро (отдельное право на скачивание файлов, загруженных в GeoNetwork)	ядро	нет

Возможность	Dataverse	InvenioRDM	DSpace	GeoNetwork	NADA	PxWeb
Роли	ядро	ядро	ядро	ядро	ядро	не проверено
Проверка перед публикацией	ядро	ядро	ядро	ядро (по умолчанию одобрение можно обойти)	не проверено	не проверено
Единый вход	ядро (SAML, OpenID Connect, ORCID)	ядро (ORCID, OpenID Connect, SAML)	ядро	ядро	не проверено	не проверено
Сбор чужих каталогов	ядро (клиент OAI-PMH; карточка ведёт на оригинал)	не проверено	не проверено	ядро (CSW, OAI-PMH, сервисы OGC, STAC)	не проверено	не проверено
Свои расширения	внешние инструменты и просмотрщики	модули платформы	расширения	плагины схем; OGC API Records — расширение	мало	мало

udata, Piveau и GeoNode

Эти три платформы не попали в таблицы выше, но в главах о возможностях уже сверены. Клетка «не проверено» здесь значит, что в тех главах ответа не было.

Возможность	udata	Piveau	GeoNode 5.1
DCAT	ядро: по страницам и файлом у набора	ядро: несколько форматов и SPARQL	нет: каталог отдаёт API и CSW, не DCAT
schema.org	нет: в HTML страницы разметки нет	нет: страница набора на data.europa.eu рисуется скриптом	нет
Каталожный протокол	ядро (DCAT); ночная выгрузка всего каталога в CSV — задача ядра	ядро (DCAT и SPARQL)	ядро (CSW и ISO 19115, внутри pycsw)
Роли	ядро: администратор, редактор, частичный редактор	не проверено	не проверено
Проверка перед публикацией	черновик — ядро; проверки куратором нет	не проверено	не проверено
Сбор чужих каталогов	ядро (DCAT, CSW, CKAN, DKAN и другие; по умолчанию выключены, после сбора пересобирают поисковый индекс)	ядро	не проверено

Чего таблица не заменяет

Клетка «ядро» не значит «включается галочкой за час». DOI в GeoNetwork и Dataverse требует договора с DataCite. Клетка «расширение» не значит «не берите»: ckanext-dcat — обычная часть взрослого портала на CKAN, а не экзотика. Клетка «нет» значит, что эту функцию придётся закрывать другой платформой или своей разработкой.

Источники

- Карточки платформ

- ckanext-dcat, возможности Dataverse
- Сбор: сборщики Huwise, проверено 14 октября 2026 года; Catalog Connector у Socrata. Наборы, подключённые к Socrata через `data.json`, после синхронизации становятся открытыми. Клиент OAI-PMH у Dataverse и шесть сборщиков udata — глава о других возможностях, сверка 14 октября 2026 года
- Роли, проверка перед публикацией и единый вход — глава о доступе и ролях, сверка 12 октября 2026 года
- Карточки программного обеспечения реестра dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года. Пометки вроде «только плагином» использовались как подсказка, что проверять в документации, а не как единственный источник.

Чек-листы

Короткие списки живут в главах, рядом с объяснением. Здесь они собраны в рабочем виде: их можно распечатать или скопировать в задачу. У каждого пункта ссылка на главу, где написано, зачем он нужен.

Два длинных списка остаются в своих главах, потому что без пояснений их не заполнить:

- Требования к каталогу и сравнение платформ — что писать в задании и что из этого обязательно не всем.
- Самоаудит — как за один проход оценить свой каталог и что записать в итог.

Проверка живого портала

Этот список подходит и для своего каталога, и для чужого, который вы рассматриваете как образец или как источник для сбора. На него уходит около получаса.

Находимость и выгрузка

- [ ] На странице набора есть блок `application/ld+json` с типом Dataset — экспорт метаданных.
- [ ] Выгрузка заявляет профиль, который ждёт сборщик (DCAT-AP 3.0.1, DCAT-US 3.0), а не просто открывается по адресу `data.json`.
- [ ] Отвечает карта сайта `sitemap.xml` — API и массовая выгрузка.
- [ ] Есть файл со всеми описаниями (`data.json`, `catalog.rdf`, `catalog.ttl`), и он открывается без входа.
- [ ] Число наборов в выгрузке совпадает с числом в поиске на сайте.
- [ ] Выгрузка проходит валидатор нужного профиля: DCAT-AP, INSPIRE или DataCite — метрики качества.
- [ ] Для геопортала отвечает CSW или OGC API Records, для научного репозитория — OAI-PMH.
- [ ] Постраничная выгрузка доходит до последней страницы, а поиск через API отдаёт следующую страницу — API и массовая выгрузка.
- [ ] Есть выборка «изменено после даты» или лента изменений, и за последние дни в ней что-то находится.
- [ ] Описания и API отвечают из командной строки, без входа, ключа и проверки от роботов, лучше из другой сети.
- [ ] Для больших баз есть массовая выгрузка файлами с датой и размером, а не только API с лимитами.
- [ ] Лимиты, условия использования и документация API опубликованы; у API есть контакт.

Описания

- [ ] У десяти случайных наборов, не самых свежих, заполнены лицензия, издатель, дата обновления, период и территория — полнота метаданных.
- [ ] Самое частое значение периода, территории и частоты обновления — не заглушка вроде «1900–2099» или «вся страна» у наборов разных организаций — проверка на выборке.
- [ ] Описание не повторяет название и длиннее 50 символов: короче не примет разметка для Google.
- [ ] Лицензию ищут там, где её хранит схема: в DCAT-AP — у файлов, а не у набора.
- [ ] Тема, лицензия, язык и территория взяты из общего справочника, а не свободным текстом — стандарты метаданных.
- [ ] У табличных наборов есть словарь полей, и типы колонок совпадают с содержимым.
- [ ] Свои поля и плагины схем (ckanext-scheming, ckanext-validation, профили GeoNetwork) проверены по репозиторию и ветке под текущую версию ядра.

Содержимое

- [ ] Большинство наборов — не справки на одну-две строки, а общий объём данных сопоставим с главными источниками темы — проверка по существу.
- [ ] «Обновлённые» наборы действительно изменились: в файле новые периоды, а не только новая дата в карточке.
- [ ] У CSV указаны или однозначны кодировка и разделитель — антипаттерны.

Файлы и адреса

- [ ] Ссылки «Скачать» открываются. Посчитайте долю внешних ссылок, которые отвечают ошибкой, и сведите ошибки по доменам — внутреннее хранилище.
- [ ] По ссылке с форматом CSV, XLSX или PDF открывается файл, а не веб-страница или главная страница сайта.
- [ ] Внешние ссылки проверяются по расписанию, а у ресурса видны дата и итог последней проверки.
- [ ] Предел размера загрузки поднят до реального: в CKAN по умолчанию 10 МиБ.
- [ ] Опубликованный файл, на который ссылаются, не перезаписывается: исправление — новая версия; у файла есть контрольная сумма.
- [ ] Пользователю дают постоянный адрес файла на портале, а не временную ссылку облачного хранилища.
- [ ] Адрес страницы набора не меняется при правке описания и названия — идентификаторы.
- [ ] Если обещан DOI, `https://doi.org/…` ведёт на страницу набора, а не на главную, тестовый стенд или `localhost`.
- [ ] Если на странице Handle, он разрешается через `hdl.handle.net`, а префикс не заглушка `123456789` из поставки DSpace.
- [ ] Снятый набор с DOI заменён страницей-надгробием, а не страницей «не найдено».
- [ ] Если портал переезжал, старые адреса перенаправляют, а не отвечают «не найдено».

Цитирование и доступ

- [ ] Есть блок «Как цитировать», скачивается BibTeX или RIS с тем же идентификатором — экспорт цитирования.
- [ ] В ссылке год публикации набора, а не год создания карточки; есть версия или, у обновляемых наборов, дата обращения.
- [ ] Если у набора есть DOI, ссылка от DataCite совпадает со ссылкой на странице по автору, году и версии.
- [ ] У набора видна организация и контакт — доступ, роли и процессы.
- [ ] Описания закрытых наборов видны без входа, скачивание — нет.

Поиск, языки, сбор и обратная связь

- [ ] Слово из середины описания находится поиском, а при сотнях наборов есть фильтр по организации и формату — другие возможности.
- [ ] После смены языка то же слово на этом языке находит набор, и перевод описания совпадает с текстом на языке оригинала.
- [ ] У табличного файла на странице есть таблица или прямая ссылка. У закрытого файла содержимое не видно тому, кому файл закрыт.
- [ ] У собранного набора есть ссылка на оригинал, а у источника сбора — дата последнего успешного прохода и число наборов.
- [ ] Счётчик видит тот способ доступа, которым данными пользуются: страницу, файл или API.
- [ ] Сообщение об ошибке доходит до владельца данных. Назван человек и срок ответа.

Сетевую часть можно проверить одним скриптом. Подставьте адрес портала:

```
PORTAL=https://portal.example
for path in sitemap.xml data.json catalog.rdf catalog.ttl \
            "api/3/action/package_search?rows=1" \
            "oai?verb=Identify" \
            "csw?service=CSW&request=GetCapabilities"; do
  code=$(curl -s -o /dev/null -w "%{http_code}" -L --max-time 20 "$PORTAL/$path")
  echo "$code  $path"
done
```

Ответ `200` по адресу, который подходит к типу каталога, — хороший знак. Ответ `401`, `403` или страница входа значит, что машинного доступа нет, даже если он обещан. Адреса для CKAN, OAI-PMH и CSW подходят не всем платформам: отсутствие ответа по чужому типу — не ошибка.

Каталог показателей

Подробно — в главе о каталогах показателей.

- [ ] Территории, отрасли и другие разрезы закодированы общими справочниками, справочники скачиваются.

- [ ] Всю базу можно забрать целиком: архивом или файлом на показатель плюс перечень файлов.
- [ ] Выгрузка из конструктора отдаёт весь ряд, а не только срез, показанный по умолчанию.
- [ ] Есть API, у каждого показателя — постоянный адрес запроса.
- [ ] Кроме Excel есть CSV и SDMX или JSON-stat.
- [ ] Методика и единица измерения лежат рядом с рядом, а не только в PDF.
- [ ] API описан по OpenAPI или хотя бы веб-страницей с примерами запросов.

Каталог метаданных

Подробно — в главе о каталогах метаданных.

- [ ] У каждого понятия, кода и описания постоянный адрес, который не меняется при переезде сайта.
- [ ] Устаревшая запись помечена, и видно, чем её заменили; молча записи не удаляются.
- [ ] Адрес записи отдаёт программе SKOS, SDMX или DCAT, а не только страницу.
- [ ] Весь словарь или перечень скачивается одним файлом.
- [ ] У каждой записи назван хранитель, у каталога записаны правила включения.
- [ ] Каталог зарегистрирован уровнем выше: в BARTOC, re3data, FAIRsharing, индексе FAIR Data Point или национальной единой точке.
- [ ] Программам не отвечает проверка от роботов вместо данных.

Каталог микроданных

Подробно — в главе о каталогах микроданных.

- [ ] Описание исследования видно без входа: выборка, единица наблюдения, период, территория.
- [ ] У переменных есть текст вопроса, значения с подписями и частоты.
- [ ] У каждого файла указан тип доступа; условия, что нужно для заявки и срок ответа написаны до заявки.
- [ ] Описания выгружаются в DDI, есть API или OAI-PMH, и они отвечают программе без проверки «я не робот».
- [ ] У исследования есть DOI или постоянный адрес и готовая ссылка для цитирования.
- [ ] Открытые файлы лежат и в CSV или Parquet, а не только в форматах SPSS и Stata.
- [ ] Каталог стоит на домене владельца с HTTPS, а не на IP-адресе или поддомене подрядчика; сопровождение заложено в бюджет.

Перед установкой расширения

Подробно — в главе об экосистемах расширений.

- [ ] Понятно, кто выпускает расширение и отвечает на ошибки. Один человек без ответа два года и без тестов — повод искать другой модуль.
- [ ] Ваша версия ядра названа в автоматических тестах, таблице совместимости или имени ветки. Свежий выпуск при старых тестах не считается. «Работало у соседей» не считается.
- [ ] Если копий несколько, в карточке решения записан адрес той, которую ставите.
- [ ] Команда установки указывает на этот репозиторий и на ветку, которая в нём есть.
- [ ] Есть файл лицензии, и она разрешает то, что вам нужно.
- [ ] Для внешней программы записано, как она получает файл: короткая подписанная ссылка или долгоживущий ключ.
- [ ] В карточке решения записано, что делать, если расширение перестанут поддерживать — как обосновать решение.

Перед покупкой или установкой платформы

Подробно — в главах о развёртывании и сопровождении и о цене ухода.

- [ ] Есть официальный контейнер и короткая инструкция обновления.
- [ ] Известно, сколько последних версий поддерживается и как часто они выходят.
- [ ] Посчитано, сколько расширений нужно для обязательных функций и кто их выпускает.
- [ ] Названо, кто будет дежурить через два года: своя ставка, подрядчик или поставщик сервиса.
- [ ] Проверено, что все описания и файлы можно выгрузить целиком, — иначе уйти с платформы будет дорого.
- [ ] Пилот прошёл на десяти настоящих наборах, а не на демонстрационных.

Перед тем как назвать набор опубликованным

Подробно — в главе о процессе публикации.

- [ ] У набора постоянный адрес.
- [ ] Лицензия на карточке, у файла и в выгрузке — одна и та же ссылка или известный код. «Не указана», код без текста и фраза в подвале не считаются — лицензии.
- [ ] Файл или сервис открывается, или честно сказано, почему он недоступен и как его запросить.
- [ ] Указаны дата обновления и периодичность.
- [ ] Описание попало в общую выгрузку и видно верхнему порталу.
- [ ] Черновик не попал в поиск до публикации.
- [ ] Назван человек или команда, которые ответят на ошибку в строках. Ящик `info@` без имени этого не закрывает.
- [ ] В перечне или очереди записано, почему этот набор публикуют сейчас.

Пока набор остаётся опубликованным

Подробно — в главе о процессе публикации.

- [ ] Дата файла не старше обещанной частоты. Пустая частота значит, что обновление для этого набора не задано.
- [ ] Снятый набор помечен: прежний адрес открывается и объясняет, почему выпуска больше нет или чем его заменили.
- [ ] Просьба о недостающем наборе попадает в очередь, а не остаётся только письмом.
- [ ] Перечень данных существует как документ — таблица или файл описаний, — а не только как память редактора.

Перед выключением каталога

Подробно — в главах о жизненном цикле и об архивации.

- [ ] О закрытии объявлено заранее, за месяцы, а не за дни; на переходный срок каталог работает только на чтение.
- [ ] Данные переданы в другой каталог или архив, либо явно сказано, что они сняты с публикации.
- [ ] Старые адреса либо перенаправляют, либо хотя бы год отвечают, куда переехал набор.
- [ ] Снимок описаний сохранён: выгрузка через API или файл DCAT, а не только страницы сайта.
- [ ] Файлы, которые лежат только в этом каталоге, сохранены вместе с адресами, откуда они взяты.
- [ ] Проверено, что сайт есть в Интернет-архиве или национальном веб-архиве и что архив до него доходит, в том числе из-за рубежа.
- [ ] Верхний портал и реестры, где каталог был зарегистрирован, уведомлены.

Перед тем как заводить свой портал

Подробно — в главе «Когда портал не нужен».

- [ ] Назван адресат: кто забирает описания и в каком профиле.
- [ ] Посчитаны наборы, организации-публикаторы и частота обновления.
- [ ] Назван человек, который будет обновлять публикацию через три года, и отдельно — кто обновляет сервер, если портал всё-таки ставят.
- [ ] Проверен чужой каталог, который уже принимает эти наборы: региональный, национальный, университетский, Zenodo.
- [ ] Если остаётся свой сайт или репозиторий, у каждого набора есть лицензия, дата и прямая ссылка на файл.
- [ ] Записано событие, при котором отказ от портала открывают заново.

Перед входом для программ и пространством данных

Подробно — в главе об ИИ и пространствах данных.

- [ ] Выгрузка описаний, лицензия и прямая ссылка на файл уже работают.
- [ ] Сервер MCP или файл `llms.txt` заводят, когда программами уже пользуются и есть кому его сопровождать.
- [ ] Поиск фразой стоит рядом с обычным поиском и возвращает карточки наборов.
- [ ] Чат ссылается на набор. Карточка, файл и лицензия остаются.
- [ ] Пространство данных заводят, когда данными обмениваются несколько организаций и права у них разные.

Перед тем как чинить качество

Подробно — в главе об уровнях качества.

- [ ] В задаче назван уровень: данные, описания, каталог как сервис или система каталогов.
- [ ] Исполнитель — тот, кто этот уровень держит: владелец данных, редактор, администратор платформы или координатор страны.
- [ ] Очередь начинается с файлов, которые открываются и читаются. Выгрузка обгоняет её, когда сверху названа дата сбора.
- [ ] Чужой балл разложен по уровню: балл карточки и MQA — описания, индекс страны — система, число наборов — счётчик карточек.

Минимум публикации без портала

Подробно — в главе о том, как сделать публикацию находимой.

- [ ] Одна страница с постоянным адресом. Путь не меняется, когда правят название.
- [ ] Описание в разметке — от 50 до 5000 знаков, тот же смысл, что в тексте страницы.
- [ ] Лицензия — ссылка на конкретную версию, одна и та же на странице, в файле и в записи DOI.
- [ ] В HTML есть schema.org/Dataset и прямая ссылка на файл. Пустого DOI в разметке нет.
- [ ] Страница открывается без входа. Если страниц несколько, есть карта сайта. `robots.txt` лежит в корне имени сайта.
- [ ] Файл в профиле сборщика — только если сборщик есть, и обязательные поля этого профиля заполнены.
- [ ] У таблицы есть `datapackage.json`. Дата обновления и контакт написаны на странице.
- [ ] DOI — если на набор ссылаются. В `sameAs` стоит другая страница того же набора.

Готовые заготовки файлов — README, `datapackage.json`, `CITATION.cff` и разметка schema.org — лежат в шаблонах.

Статический каталог

Подробно — в главе о статических сайтах.

- [ ] Страница каждого набора собрана в HTML. Текст описания есть в файле страницы, а не подгружается только скриптом.
- [ ] Адрес страницы не меняется, когда правят название. Переименование файла адрес меняет.
- [ ] В HTML есть schema.org/Dataset. Разметка DCAT в атрибутах страницы её не заменяет.
- [ ] Файл всех описаний открывается без входа. В записи есть издатель, лицензия, дата изменения и прямая ссылка. Строка `conformsTo` сверена с полями.
- [ ] В поле лицензии стоит известная лицензия. Значение `No licence` не считается.
- [ ] Карта сайта `sitemap.xml` открывается.
- [ ] Если включена форма редактора, правка видна как коммит. Отдельный сервер входа, если хостинг его требует, записан в карточке решения.
- [ ] Размер сайта и время сборки укладываются в предел хостинга.

Репозиторий с данными

Подробно — в главе о Git-репозиториях.

- [ ] Один набор — один пакет. Описатель называется `datapackage.json` и лежит в корне пакета.
- [ ] В `LICENSE` — лицензия на данные. Если рядом скрипт, лицензия кода названа отдельно.
- [ ] В `CITATION.cff` есть `type: dataset`. Пустого DOI в файле нет.
- [ ] Файл меньше предела площадки. Большой файл — во вложении выпуска или вне Git, в репозитории остаётся адрес.
- [ ] Выпуск, на который ссылаются, зафиксирован меткой. На GitHub связка с Zenodo включена до этого выпуска, у записи тип «набор данных», лицензия совпадает с `LICENSE`.

Файлы в облаке или на общем хабе

Подробно — в главе об облачных хранилищах и хабах.

- [ ] Рядом с файлами лежат `README.md`, лицензия и контрольные суммы. Лицензия на странице, в карточке и в файле одна и та же.
- [ ] Адрес файла постоянный. Это не временная ссылка и не бакет с оплатой трафика читателем.
- [ ] Файл скачивается без входа. Проверено запросом без ключа облака и в браузере, где вы не вошли на площадку.
- [ ] У выпуска, на который ссылаются, есть DOI или копия вне площадки.
- [ ] До загрузки проверено, что свои наборы можно забрать целиком. Если место оплачивает программа спонсорства, в карточке решения записан день, когда срок кончится.

Литература и ссылки

Сюда собраны внешние источники из разделов «Источники» глав. Ссылка на соседнюю главу книги в этот список не входит.

Реестры каталогов

- dataportals-registry, выпуск 1.27.0
- Оценка доверия в реестре
- re3data
- FAIRsharing
- dataportals.org, Open Data Inception, каталог экосистемы CKAN
- Реестры словарей и онтологий: BARTOC, Linked Open Vocabularies, Ontology Lookup Service, индекс FAIR Data Point
- Европейский реестр охраняемых данных государственного сектора на data.europa.eu
- Registry of Open Data on AWS и программа спонсорства открытых данных AWS
- Source Cooperative

Стандарты

- DCAT 3 (W3C), рекомендация от 22 августа 2024 года
- DCAT-AP 3.0.1, рекомендация SEMIC от 27 октября 2025 года; черновик 3.0.2
- DCAT-US 3.0
- GeoDCAT-AP 3.1.0
- schema.org/Dataset
- SKOS, XKOS, OWL 2
- ISO/IEC 11179-3:2023, реестры метаданных
- HealthDCAT-AP, проект спецификации 2026 года
- Спецификация FAIR Data Point
- DataCite и схема метаданных DataCite 4.7
- Постоянные идентификаторы: состояния DOI в DataCite, страницы-надгробия, взносы DataCite, санкционная политика Crossref, Handle.Net Registry, ARK Alliance, w3id.org, PURL, CSTR
- DDI
- SDMX
- STAC
- OAI-PMH 2.0, robots.txt, RFC 9309
- Frictionless Data Package и Table Schema, CSV on the Web, каталог схем schema.data.gouv.fr
- CITATION.cff

- Цитирование: согласование содержимого DataCite, репозиторий стилей CSL, как отдать метаданные Zotero, ГОСТ Р 7.0.108-2022 — ссылки на электронные документы в сетях
- Croissant 1.1
- RO-Crate 1.3, CDIF
- Датасеты Hugging Face

Платформы

- Документация CKAN 2.12 и каталог экосистемы CKAN — расширения и порталы
- Порядок доработок InvenioRDM, внешние инструменты Dataverse, плагины схем GeoNetwork, модули сообщества GeoServer
- ckanext-dcat
- ckanext-doi, ckanext-citeproc
- Свои поля описания: ckanext-scheming, блоки метаданных Dataverse, свои поля InvenioRDM, плагины схем GeoNetwork, шаблоны Huwise, редактор метаданных Всемирного банка
- Возможности Dataverse и цитирование данных
- Лицензия InvenioRDM, MIT
- Лицензия GeoNetwork, GPL-2.0
- Лицензия DKAN, GPL-2.0
- PxWeb и PxWebApi 2, .Stat Suite, SDMX-RI, Fusion Metadata Registry, DHIS2
- SDMX, JSON-stat, StatDCAT-AP
- ERDDAP, wis2box, Fairdata
- Сбор карточек в GeoNetwork, pycsw
- JKAN, Open SDG, PortalJS, Datasette
- Zenodo и архивация выпусков GitHub
- Каталоги метаданных: FAIR Data Point, Skosmos, OntoPortal, OLS, MOLGENIS

Практики и методики

- Michael Nygard, Documenting Architecture Decisions, 15 ноября 2011 года — короткая запись решения: контекст, само решение, последствия; заменённую запись оставляют
- Data on the Web Best Practices (W3C), 31 января 2017 года — частота обновления, объяснение недоступных данных, сбор и публикация отзывов
- GSA, DCAT-US v3.0 Implementation Guide, редакция 1.1 от 9 сентября 2026 года — инвентаризация, обязательный контакт, снятие набора с объяснением
- Руководство data.gouv.fr: опубликовать набор — при регулярном обновлении карточка уходит через API
- Принципы FAIR
- Международная хартия открытых данных, 2015
- 5 звёзд открытых данных

- T. E. Hardwicke и др., Analytic reproducibility in articles receiving open data badges at the journal Psychological Science, Royal Society Open Science, 2021 — из 25 статей со значком открытых данных без помощи авторов проверенные числа сошлись у 9
- Принципы цитирования данных, FORCE11, 2014
- Research Data Alliance, Precise, Actionable Reference to Dynamic Data, 2025 — как ссылаться на меняющиеся данные и их части
- Make Data Count, Data Citation Corpus — корпус цитирований данных; правила цитирования GBIF — DOI выгрузок и производные наборы
- Creative Commons
- Open Definition 2.1 — что считать открытой лицензией: коммерческое использование разрешено, запрет коммерции и запрет изменений — нет
- Список кодов SPDX
- CoreTrustSeal
- Open Data Maturity, отчёт 2025 года — опрос 36 стран; средний балл ЕС-27 86%, всех участников 81%; методика — четыре измерения по 25%
- MQA — оценка карточек, которые data.europa.eu забрал у порталов; шкала каталога от 0 до 7,5; методика, страница от 6 октября 2026 года
- ОЭСР, Digital Government Index and OURdata Index: 2025 Results and Key Findings, 16 февраля 2026 года — средний балл стран ОЭСР 0,53 против 0,48 в выпуске 2023 года
- Global Data Barometer, второе издание, 29 мая 2025 года — 43 страны Африки и Латинской Америки
- Open Data Inventory (ODIN), выпуск 2024/25 — 197 стран
- Open Data Barometer — последнее полное издание в мае 2017 года; Global Open Data Index — архив, последнее издание 2016/17

Как ищут данные

- N. W. Paton, J. Chen, Z. Wu. Dataset Discovery and Exploration: A Survey. ACM Computing Surveys, 2023 — обзор методов поиска и исследования наборов данных
- K. Gregory, P. Groth, A. Scharnhorst, S. Wyatt. Lost or Found? Discovering Data Needed for Research. Harvard Data Science Review, 2020 — как исследователи ищут чужие данные, по результатам международного опроса
- K. Sostek, D. M. Russell, N. Goyal, T. Alrashed, S. Dugall, N. Noy. Discovering Datasets on the Web Scale: Challenges and Recommendations for Google Dataset Search. Harvard Data Science Review, 2024 — опыт Google Dataset Search
- Web Data Commons: разметка schema.org из Common Crawl — сколько и какой разметки Dataset встречается в вебе
- Handbook on Using Administrative Data for Research and Evidence-based Policy — как найти и получить административные данные для исследований
- Блог И. Бегтина begtin.substack.com — разборы порталов, поиска данных и практик публикации, на которые ссылаются главы книги
- Telegram-канал И. Бегтина @begtin — короткие заметки о порталах, их закрытии и архивации с 2016 года; на отдельные посты ссылаются разделы «Источники» глав

Сохранность и закрытие данных

- Pew Research Center, When Online Content Disappears, 2024 — доля исчезнувших страниц и битых ссылок на сайтах
- T. Berners-Lee, Cool URIs don't change, 1998; политика постоянства адресов W3C
- M. Laakso, L. Matthias, N. Jahn, Open is not forever: a study of vanished open access journals, 2020
- Sunlight Foundation, Classifying changes to public access to information on US government websites, 2017 — восемь способов исчезновения государственной информации
- «Если быть точным», В 2025 году в России исчезли 140 датасетов и перестали обновляться 425 показателей, 2025
- Congressional Research Service, Data.gov: Implementation and Perspectives on Its Functions, 2026
- End of Term Web Archive, Data Rescue Project, Harvard LIL: архив data.gov, Национальный цифровой архив
- Закрытие площадок: data.world, A fond farewell, 2026; Microsoft, Retiring the Planetary Computer Hub, 2024; QRI, Winding Qri Down, 2022
- BagIt, RFC 8493, WACZ

Данные для ИИ

- Open Data Institute, The UK government as a data provider for AI, ноябрь 2024 — общие модели почти не помнят цифры data.gov.uk
- Open Data Institute, A framework for AI-ready data, версия 2, декабрь 2025
- Open Data Product Specification — описание дата-продукта
- Data Spaces Support Centre, Blueprint 3.0, февраль 2026; CEN/CLC/TS 18331:2026 — оценка зрелости пространства данных
- OMB, меморандум M-25-05, 15 января 2025 — инвентаризация всех активов данных, обновление не позднее 90 дней и не реже раза в год, файл `data.json`; среди целей открытого доступа названо обучение ИИ
- Open Knowledge Foundation, MCP for Open Data Portals: Trust Depends on Understanding the Data, июнь 2026 — пилоты MCP в Бразилии и Уругвае, частые вопросы как тесты
- Open Data Watch, Making Official Statistics Work in an AI-Driven World, сентябрь 2026 — официальные серверы MCP статистических служб
- Всемирный банк, Data360 MCP — сервер с инструкцией для агента и проверкой покрытия

Передовые каталоги

- data.gouv.fr, API tabulaire — API к каждому табличному файлу
- Overture Maps, GERS — постоянные идентификаторы объектов
- DoltHub, What is Dolt? — версии таблиц как в Git

- OpenNeuro, руководство пользователя; DANDI, публикация наборов — проверка на входе
- Copernicus Data Space Ecosystem, openEO — обработка рядом с данными
- Data Commons, HDX HAPI, OpenSanctions, модель утверждений — сведение многих источников
- microdata.no, Redivis, уровни доступа — доступ по ступеням

Право ЕС

- Директива INSPIRE 2007/2/EC
- Директива (ЕС) 2019/1024 об открытых данных
- Регламент (ЕС) 2023/138 о наборах высокой ценности
- Регламент (ЕС) 2022/868, Data Governance Act
- Регламент (ЕС) 2025/327 о Европейском пространстве медицинских данных (EHDS)
- Регламент (ЕС) 2023/2854, Data Act
- Регламент (ЕС) 2024/1689 об искусственном интеллекте — сводка обучающих данных у поставщика модели общего назначения

Стабильные адреса для ссылок

Постоянный адрес книги — datacatalogs.begtin.tech. На главы и разделы можно ссылаться извне: их адреса не меняются молча.

Адреса книги

Версия	Сайт	PDF
Русская	datacatalogs.begtin.tech	data-catalogs-book.pdf
Английская	datacatalogs.begtin.tech/en/	data-catalogs-book-en.pdf

Адрес главы на английском — тот же путь с приставкой `/en/`: например, `https://datacatalogs.begtin.tech/types/classification/` и `https://datacatalogs.begtin.tech/en/types/classification/`. Полный адрес главы — домен и путь со слэшем в конце.

Прежний адрес книги `ivbeg.github.io/data-catalogs-book/` перенаправляет на новый домен вместе с путём главы, поэтому старые ссылки продолжают работать. Как сослаться на книгу целиком, сказано в разделе «Об издании».

Как адреса остаются стабильными

Путь главы не меняют молча. Если главу переименуют, со старого адреса настраивают перенаправление. Старые пути про API и идентификаторы уже ведут на новые главы.

Список путей для базы знаний реестра ведут вместе с сайтом, отдельно от текста для читателя.

Об авторе



Иван Бегтин занимается открытыми данными: собирает и разбирает данные государства и науки, делает сервисы, которые помогают их найти, и пишет о том, как их публикуют. Он основатель и технический директор поисковика по наборам данных Dateno, основатель и директор АНО «Информационная культура» (Инфокультура) и сооснователь проекта «Открытые данные Армении» (opendata.am).

Открытостью государства и открытыми данными он занимается с 2009 года, каталогами данных — с 2010-го. Эта книга выросла из той же работы. Реестр каталогов, по которому посчитаны её цифры, он начал для своих проектов, а многие примеры и подсчёты в главах взяты из его заметок.

Путь

В ИТ Иван Бегтин работает с 1999 года. Начинал с администрирования серверов Unix, потом был разработчиком и руководил лабораторией интеллектуального анализа данных.

В 2009 году он запустил проект «Государство и его информация» (opengovdata.ru) — новости и базу знаний об открытых государственных данных в России. Проект проработал до 2015 года. Тогда же он начал сохранять копии сайтов органов власти, которые закрывались или менялись.

В 2010 году он стал одним из учредителей конкурса Apps4Russia, где разработчики делали приложения на государственных данных. В марте 2012 года основатели конкурса создали АНО «Информационная культура», и с тех пор Иван Бегтин ею руководит. Среди проектов Инфокультуры — открытая база государственных контрактов «Госзатраты» (clearspending.ru) и ежегодный День открытых данных в Москве.

В 2019–2021 годах он руководил проектным направлением «Открытые данные» в Счётной палате России. Команда из шести человек за четыре с половиной месяца сделала там портал «Госрасходы» (spending.gov.ru). Его запустили в ноябре 2019 года: на портале были контракты, субсидии и расходы на национальные проекты, а сами сведения отдавались открытыми данными и через API.

Ещё одна давняя тема автора — понятный язык в текстах государства. В 2014 году он адаптировал к русскому языку формулы удобочитаемости, которыми в англоязычных странах проверяют официальные тексты. Эта книга написана по тем же правилам: простые слова и сначала вывод.

Каталоги данных

Первым каталогом стал «Хаб открытых данных» hubofdata.ru на CKAN. Иван Бегтин сделал его в 2010–2011 годах под маркой OpenGovData, и он появился раньше государственных порталов открытых данных в России (об этом — в короткой истории каталогов). Позже Инфокультура запустила на той же платформе портал данных для некоммерческих организаций «Данные НКО» ngodata.ru. Оба общественных портала книга вспоминает в главе об экспорте метаданных: в ноябре 2024 года Google Dataset Search по русским запросам выдавал их наборы на первых местах, потому что на их страницах есть разметка schema.org для поисковиков.

Через десять лет каталогов стало так много, что понадобились каталоги каталогов. В ноябре 2021 года Инфокультура запустила перечень каталогов данных России datacatalogs.ru: при запуске в нём было 245 записей. В 2023 году работа вышла за пределы одной страны. В проекте Common Data Index начался открытый реестр каталогов всего мира, в апреле 2023 года в нём было 1,7 тысячи записей. Теперь это реестр dataportals-registry: в выпуске 1.27.0 от 15 октября 2026 года описаны 43 614 каталогов и 790 платформ.

В марте 2024 года на основе реестра заработал поисковик Dateno. Он узнаёт из записи реестра платформу каталога и адреса его API и по ним забирает описания наборов. На 16 октября 2026 года в Dateno 22 миллиона наборов из 5 236 источников. Почему одни каталоги собрать легко, а до других сборщик не доходит, разобрано в главах «Поисковики и агрегаторы» и «Находимость».

Сохранение данных и инструменты

С 2017 года Инфокультура ведёт Национальный цифровой архив (ruarxive.org). Он сохраняет сайты и данные, которые могут исчезнуть, прежде всего государственные. В 2022 году архив провёл кампанию архивации порталов открытых данных, в 2025 году сохранил сайты Росстата и систему статистики ЕМИСС. Как устроена такая работа, рассказано в главе «Архивация и спасение каталогов».

Для своей работы автор пишет инструменты с открытым кодом. undatum обрабатывает файлы данных из командной строки: CSV, NDJSON, XML и другие форматы. metacrafter определяет, что лежит в колонках таблицы: идентификаторы, коды, персональные данные. apibackuper сохраняет данные, которые отдаются только постранично через API. Он упоминается в главах об API и массовой выгрузке и об архивации.

Автор и источники книги

Книга и база знаний о каталогах данных создаются при поддержке Dateno. Реестр каталогов и Dateno — проекты автора, поэтому книга ссылается на них чаще, чем на чужие. Это можно проверить. Цифры реестра получены открытыми запросами, они собраны в приложении «Как работать с реестром» и на выпуске 1.27.0 дают те же числа. Версии программ и возможности платформ сверены с документацией самих платформ, а не с реестром.

Заметки автора из Telegram-канала и рассылки в разделах «Источники» подписаны «И. Бегтин», у каждой стоит дата. Это наблюдения на день заметки. Если цифра с тех пор могла измениться, сверяйтесь с датой.

Где читать и как связаться

- Telegram-канал «Ivan Begtin» — открытые данные, инженерия данных, государство, приватность и сохранение цифрового наследия. На 16 октября 2026 года у канала 9 093 подписчика.
- Рассылка begtin.substack.com — длинные тексты о цифровом, открытом и сохраняемом государстве.
- Код — github.com/ivbeg, профиль в LinkedIn — linkedin.com/in/ivbeg, ORCID — 0000-0002-9084-9643.
- Личный сайт — begtin.tech: книги, проекты и тексты.
- Почта — ivan@begtin.tech.

Замечания и исправления к книге можно прислать на почту или в Telegram.

Источники

- Профили автора: GitHub — проекты Dateno, APICrafter, «Открытые данные Армении», Инфокультура; LinkedIn — «Государство и его информация» (2009–2015), сохранение сайтов органов власти (с 2009 года), Apps4Russia (с 2010 года), Национальный цифровой архив (с 2017 года); описание Telegram-канала — темы канала и число подписчиков на 16 октября 2026 года; Facebook — сооснователь Open Data Armenia, Счётная палата в 2019–2021 годах
- TAdviser, Бегтин Иван Викторович — работа в ИТ с 1999 года, лаборатория интеллектуального анализа данных, соучредитель Apps4Russia
- «Данные НКО», карточка АНО «Информационная культура» — организацию создали в марте 2012 года основатели конкурса Apps4Russia
- TAdviser, Госрасходы — запуск 18 ноября 2019 года, разработка с июня по ноябрь 2019 года силами шести человек, открытые данные и API; Росконгресс, карточка спикера — руководитель проектного направления «Открытые данные» Счётной палаты
- И. Бегтин, Что такое «Понятный русский язык» с точки зрения технологий, Хабр, 2 октября 2014 — формулы удобочитаемости для русского языка
- И. Бегтин, Telegram: Хаб открытых данных, созданный в 2010–2011 годах под маркой OpenGovData, 14.04.2021; запуск datacatalogs.ru, 03.11.2021; Common Data Index и реестр каталогов, 10.04.2023; запуск Dateno, 13.03.2024
- Реестр каталогов данных dataportals-registry, выпуск 1.27.0 от 15 октября 2026 года; Dateno на 16 октября 2026 года
- Национальный цифровой архив; «Открытые данные Армении»