Common Data Index

Как построить поисковую систему по открытым данным такую же, как Google Dataset Search, но проще и быстрее

Иван Бегтин, 2023,
Smart Data Conf

Почему я пришёл к идее создания поисковой системы по данным?

- Много лет занимаюсь открытыми данными в России и в мире
- Работаю с данными ежедневно
- Не нравятся существующие поисковые системы
- Вижу экспериментальные проекты, понимаю что можно сделать лучше

Задайте себе вопрос.
Как и где я ищу данные?

Как ищут данные?

Поиск (обычный)

- Google
- Bing
- Yandex

Поиск по данным

- Google Dataset Search
- DataCite
- OpenAIRE
- GeoSeer
- BASE
- Findata.cn

Крупнейшие каталоги

- Data.europe.eu
- Data.gov
- Zenodo
- ScienceDb
- Mendeley Data
- DataOne
- Kaggle
- Hugging Face
- ArcGIS Hub

oogle

Russian statistics

Last updated | Download format | Usage rights | Topic | Free

Saved dataset

00+ datasets found

atista Russian population size 1959-2022
statista.com
Updated Dec 14, 2022

atista Users of Meta in Russia 2022, by platform
statista.com
Updated May 13, 2022

atista Population under the poverty line in Russia quarterly...
statista.com
Updated Dec 2, 2022

E European Union Trade Balance: EU 27E: Russia: Manufactured...
ceicdata.com
Updated May 22, 2018

atista Number of Facebook users in Russia monthly 2020-2022
statista.com
amdsuportdrivers.org

European Union Trade Balance: EU 27E: Russia: Manufactured Products

Explore at: European Union | Economic I...

Dataset updated

May 22, 2018

Dataset provided by

CEICdata.com

License

Attribution 4.0 (CC BY 4.0)
License information was derived automatically

Time period covered

Sep 1, 2021 - Aug 1, 2022

Area covered

Europe, European Union

Variables measured

Merchandise Trade

Description

Trade Balance: EU 27E: Russia: Manufactured Products data was reported at 1.914 EUR bn in Sep 2022. This records an increase from the previous number of 1.669 EUR bn for Aug 2022. Trade Balance: EU 27E: Russia: Manufactured Products data is updated monthly, averaging 3.948 EUR bn from Jan 2002 to Sep 2022, with 249 observations. The data reached an all-time high of 8.132 EUR bn in Nov 2012 and a record low of -0.664 EUR Mar 2022. Trade Balance: EU 27E: Russia: Manufactured Products data remains active status in CEIC and is reported by Eurostat. The data is categorized under Global Database's European Union – Table EU.JA038: Euros Trade Statistics: By SITC: European Union: Russia.

Как устроен поиск Google

- **13 тысяч источников данных и 45 миллионов наборов данных (на 28 февраля 2023 г., блог проекта)**
- Обходит все веб-сайты, выбирает страницы с наборами данных
- В основе стандарт Schema.org тип Dataset
- Охватывает открытые и научные данные
- По умолчанию "доверяет" источникам данных
- "Захламлен" SEO оптимизаторами

Сделать поиск не хуже чем Google Dataset Search:

- только по доверенным источникам данных,
- с большим числом фасетов и фильтров
- без поискового спама
- при минимуме ресурсов
- из "подручных средств"

Изначальные предположения

Гипотезы

1. Можно создать поисковую систему аналогичную Google Search без обхода всей сети
2. Можно использовать API предоставляемое каталогами данных
3. Всё это можно сделать в относительно короткие сроки
4. На данные может не быть ссылок, поэтому не работают PageRank и другие алгоритмы
5. В основе поиска должен быть поиск по множеству фасетов
6. Можно углубить поиск и сделать поиск по полям данных и их семантическим типам

Эксперименты

Создание Datacatalogs.ru

Каталог из собранных вручную каталогов данных по России и ближайшим к России странам

Парсеры к каталогам данных

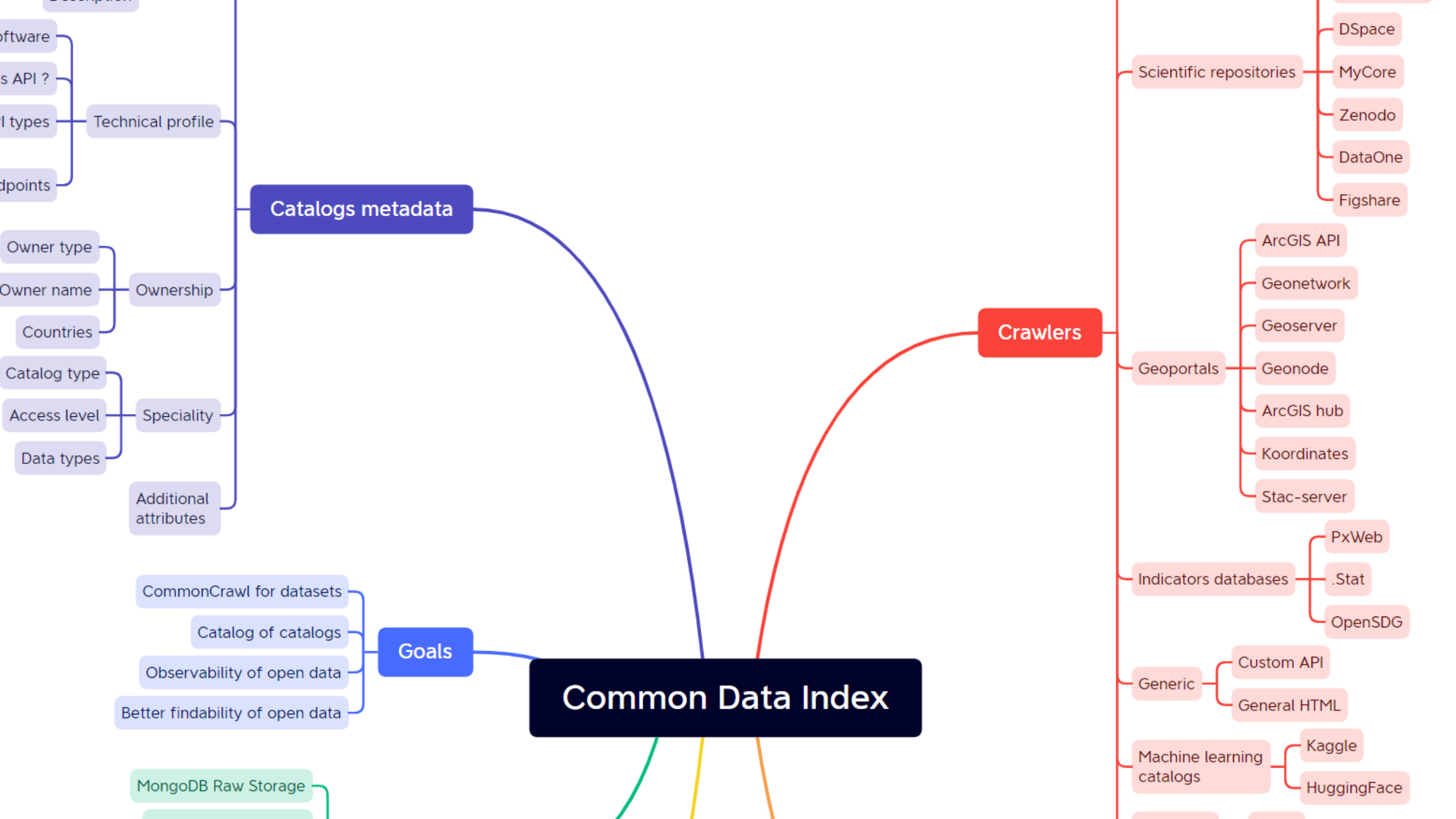
Подготовка парсеров к нескольким наиболее популярным каталогам данных таким как data.gov.uk, data.gouv.fr, data.gov.ru

Идентификация семантических типов данных

Обработка всех наборов данных из data.gov.uk, data.gov.ru и других для идентификации семантических типов данных. Всего 10 каталогов, около 100 тысяч наборов данных

Фасетный поиск

Развёртывание Elasticsearch, Meilisearch, поиска средствами Postgres, MongoDB и другими продуктами


Common Data Index
Catalogs metadata
Technical profile
ftware
s API ?
l types
dpoints
Ownership
Owner type
Owner name
Countries
Speciality
Catalog type
Access level
Data types
Additional attributes
Goals
CommonCrawl for datasets
Catalog of catalogs
Observability of open data
Better findability of open data
MongoDB Raw Storage
Crawlers
Scientific repositories
DSpace
MyCore
Zenodo
DataOne
Figshare
Geoportals
ArcGIS API
Geonetwork
Geoserver
Geonode
ArcGIS hub
Koordinates
Stac-server
Indicators databases
PxWeb
.Stat
OpenSDG
Generic
Custom API
General HTML
Machine learning catalogs
Kaggle
HuggingFace

Этапы и продукты

Реестр всех каталогов данных всех возможных типов: открытые данные, геоданные, научные данные, справочники и тд.

Первичные данные стандартизованные в единый формат схожий со стандартом DCAT

Описания наборов данных (метаданные) собранные как есть из первоисточников и доступные в виде баз данных и API

Полноценная поисковая система по всем собранным наборам данных

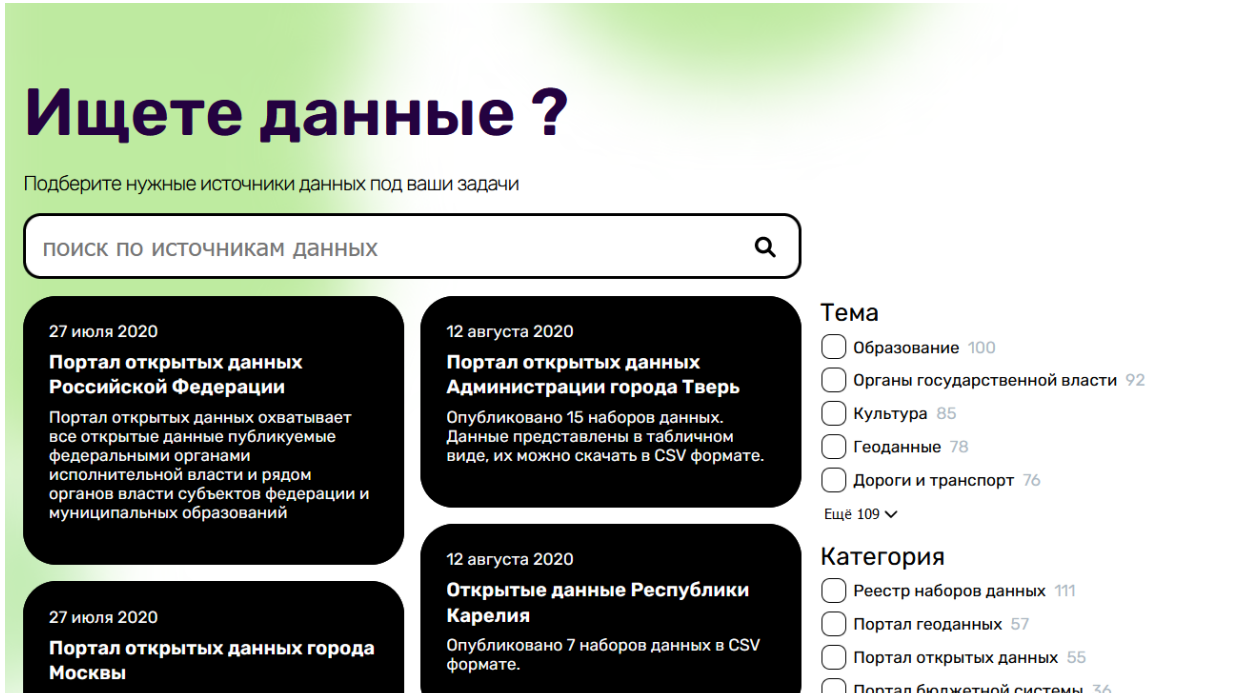
Этап 1.
Реестр каталогов данных

Ситуация

- Нет единого реестра всех каталогов данных в мире
- Есть отдельные списки каталогов данных вроде DataPortals.org, DataShades, OpenDataInception и тд.
- Во многих источниках данные очень "грязные", неактуальны, каталоги недоступны и так далее
- Хорошие метаданные есть только о каталогах научных данных
- Без каталога данных невозможно реализовать фасетный поиск поскольку многие атрибуты наборы данных наследуют от каталога данных

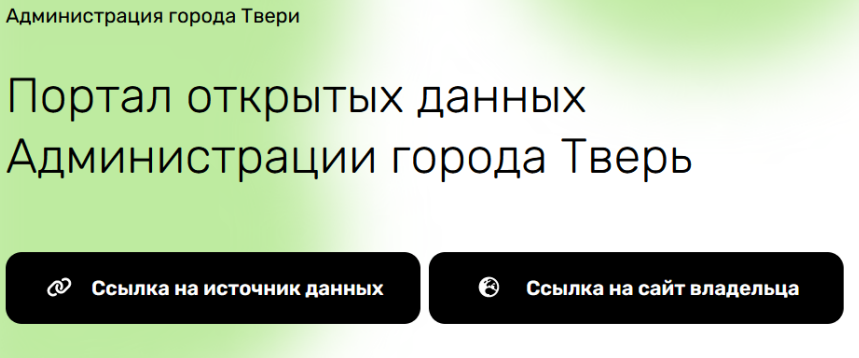
Виды каталогов данных

- Порталы открытых данных
- Репозитории научных данных
- Порталы/каталоги/сервера с геоданными
- Порталы микроданных
- Каталоги индикаторов (статистика)
- Каталоги API
- Маркетплейсы данных
- Может ещё что-то?



До реестра Common Data Index у нас был каталог каталогов данных по России, datacatalogs.ru, именно он стал основной для реестра.

Datacatalogs.ru



Реестр каталогов данных в цифрах

9751 каталогов

В основном это каталоги геоданных и порталов открытых данных. Более всего каталогов данных на базе ArcGIS Hub, CKAN и ArcGIS Server

89 типов программных продуктов

Включая такие популярные платформы как CKAN, DKAN, OpenDataSoft, Socrata, ArcGIS Hub, ArcGIS Server, Geoserver, Geonetwork, Geonode и другие

76 видов API и стандартов

Наиболее популярны DCAT, CKAN API, OAI-PMH 2.0, ArcGIS Rest, WFS, WMS, Geonetwork REST, CSW, Geonode и Dataverse.

20+ источников + ручной поиск

Около 20 основных источников для наполнения и ручной поиск данных в поисковых системах

Как создавался реестр

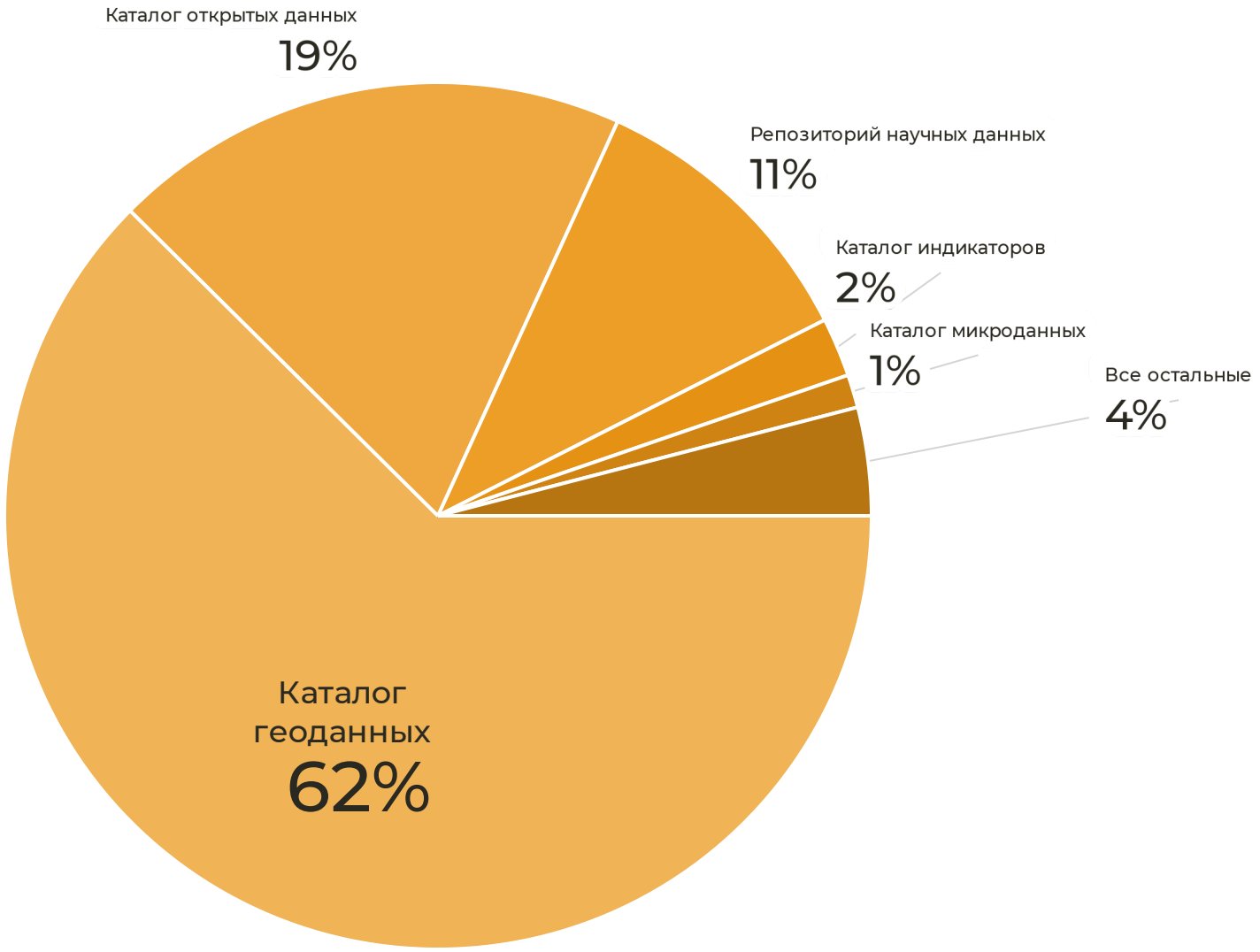
- Импорт из существующих реестров (DataShades, DataPortals.org и др.)
- Поиск в специализированных сервисах (Builtwith)
- Импорт из списков SaaS платформ (ArcGIS Hub, OpenDataSoft, Socrata)
- Ручной поиск в Google по типовым шаблонам

  По ключевым словам и частям ссылок. Например, "Powered by CKAN" или "site:ru inurl:/rest/services"

- Технологиями OSINT

  Сканирование DNS SaaS платформ, поиском префиксов доменов в Common Crawl и тд.

Статистика (всего 9751 каталогов данных)



Ограничения

- Требуется много ручной работы. Вручную вносятся:
  - тип владельца каталога данных
  - детализация до субрегионального уровня область/регион/город
  - страны для каталогов где они не указаны
  - названия и сайты владельцев каталогов данных (частично)
- Требуется регулярный мониторинг недоступных и исчезнувших каталогов

Результаты

- Самый крупный в мире реестр каталогов данных
- База знаний по ПО и API
- Фундамент для построения поисковой системы

Этап 2.
База первичных метаданных

Ситуация

- Есть много типовых API таких CKAN, Geonetwork, Dataverse, Invenio, DKAN, uData, DCAT 2.0, OAI-PMH 2.0 и других
- Большая часть API возвращают JSON, реже XML
- Авторизация почти никогда не требуется (редкие исключения)

Ограничения

- **Почти нет плоских данных**
  Простые и понятные SQL инструменты тут не работают
- **Некоторые каталоги используют авторизацию для API, но поддерживают Schema.org Dataset**
  Например, многие инсталляции Dataverse требуют авторизации для API поиска, но отдают без авторизации метаданные в API доступа к отдельным записям и на страницах датасетов.
- **Спецификации стандартов часто нарушаются (DCAT, OAI-PMH 2.0)**
  К примеру, каталоги данных на базе DKAN экспортируют файл /data.json по стандарту DCAT, но внутри записей нет ссылок на набороы данных.
- **Источник может поддерживать множество протоколов, необходимо выбирать приоритетный**
  Например, GeoNode поддерживает стандарты DCAT, WFS, WMS, а GeoNetwork одновременно собственное API, экспорт в DCAT и CSW.
- **Если добавить индекс DataCite то будет +30 миллионов наборов данных и поисковик по научным данным**
  Это важное ограничение в том что научные данные, в значительной степени, каталогизированы и если добавить индекс DataCite то всего остального останется слишком мало чтобы быть глобальным поисковиком

Конструируем из "говна и палок"

Анализ

Пришлось создавать профиль и документировать каждый программный продукт и типовые API

Docusaurus + Markdown

Парсеры

Парсеры из большого разнообразия инструментов кода на Python, конфиг файлов для утилиты APIBackuper и библиотек для работы с геопротоколами.

Очень много "так-себе-кода" на Python

Дата-инженерия?!

Пока нет шедулера задач, инкрементальных обновлений, промышленного ETL/ELT

MongoDB чтобы пока не заморачиваться схемами данных.

Огромная файловая свалка на файловой системе + MongoDB сервер

Результаты

1. 20 парсеров для типовых каталогов данных
2. около 20 парсеров в разработке
3. 1500 обработанных каталогов данных
4. 6 миллионов наборов данных

Этап 3. Поисковый индекс

Ситуация

- Разная структура для разных типов каталогов
- Огромное число каталогов данных с индивидуальными метаданными и API
- Далеко не всегда атрибуты необходимые для фасетного поиска содержатся в метаданных
- Немногие первоисточники содержать сведения о полях данных

И для обработки ещё больше "говна и палок"

И снова анализ

Заявленное соблюдение стандартов не соблюдается, схемы данных нестабильны, меняются типы полей, огромная вариативность. Приходится всё документировать

Docusaurus + Markdown

Снова парсеры

Код для преобразования из первичных данных в унифицированный формат для поискового индекса

Ещё больше "так-себе-кода" на Python

Обработка данных

Трансформация данных через преобразование на Python данных из MongoDB и загрузкой в другую таблицу в MongoDB

Уже медленно работает на миллионах записей

А есть ли вообще нормальный инструмент трансформации данных для NoSQL ?

Можно ли приспособить к этому dbt?

```
"_id": {
  "$oid": "64d1e517753604cb7af47f2c"
},
"dataset": {
  "id": "4eaf89ca-b530-4572-bc0b-e6874137c2ab",
  "url": "https://ozone.unep.org/",
  "title": "ODS Consumption in ODP,  2012 - 2017",
  "description": "<p>Data submitted to the UN Ozone
exports) in Kiribati. Ozone Depleting Substances calc
  "has_archive": false,
  "responsible": [
    {
      "title": "Environment and Conservation Divisi
      "role": "Publisher"
    }
  ],
  "topics_original": [
    "Atmosphere and Climate"
  ],
  "num_resources": 4,
  "tags": [
    "ods",
    "ozone depleting substances",
    "hcfcs",
```


Типичная запись в поисковом индексе

Поддерживаемые фасеты

	Откуда получаем	Примеры
Тип каталога	реестр	Geoportal, Open data portal
Тип владельца	реестр	Regional government, Academy
Страна	реестр	Spain, Greece, Mexico
Разговорный язык	реестр	AR, EN, ES, RU
ПО каталога данных	реестр	CKAN, DKAN, Dataverse
Каталог данных	реестр	Russian government open data portal
Регион страны / город (ISO 3166-2)	реестр	US-TX, RU-KRD
Формат	набор данных	CSV, XML, JSON
Тэг	набор данных	climate change, Europe, covid-19
Тема (открытые данные)	реестр + ML	Environment
Тема (геоданные)	реестр + ML	Biota
Лицензия	ML	CC-BY 4.0, CC-BY-SA 4.0, ODbL

[Пока] не реализованные фасеты

	Откуда получаем	Примеры
Доступность данных	реестр + набор данных	open, limited, closed
Признак геоданных	реестр + набор данных	Yes, No
Признак данных для ML	реестр + набор данных	Yes, No
Год публикации	набор данных	1900-2023
Тип поля данных	набор данных + файлы набора данных	string, int, float, datetime
Семантический тип	ML	KLADR, INN, Person name, IBAN, US tax code, DUNS
Научная дисциплина	реестр + набор данных + ML	Chemistry, Meteorology, Hydrology
Пространственный поиск	реестр + набор данных + ML	Координаты точек для поиска
Тип данных	реестр + набор данных + ML	Dataset, Map layer, Indicator

Есть ли место для ML ?

Лицензии/Права на использование

Большая вариантивность, ограниченный список

Темы/категории

Вариативность огромна, список ограничен пользовательским интерфейсом

Геопривязка

Значительные отличия в том как описывается в первоисточнике. Частично решается вручную.

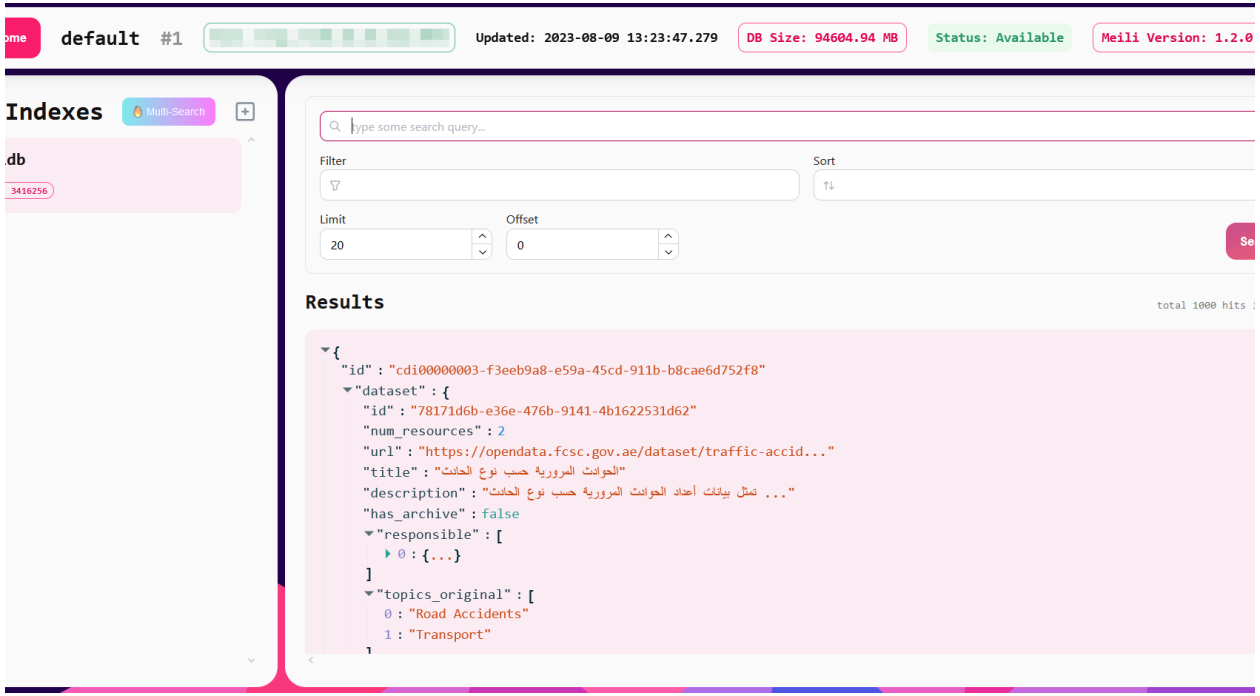
Семантические типы данных

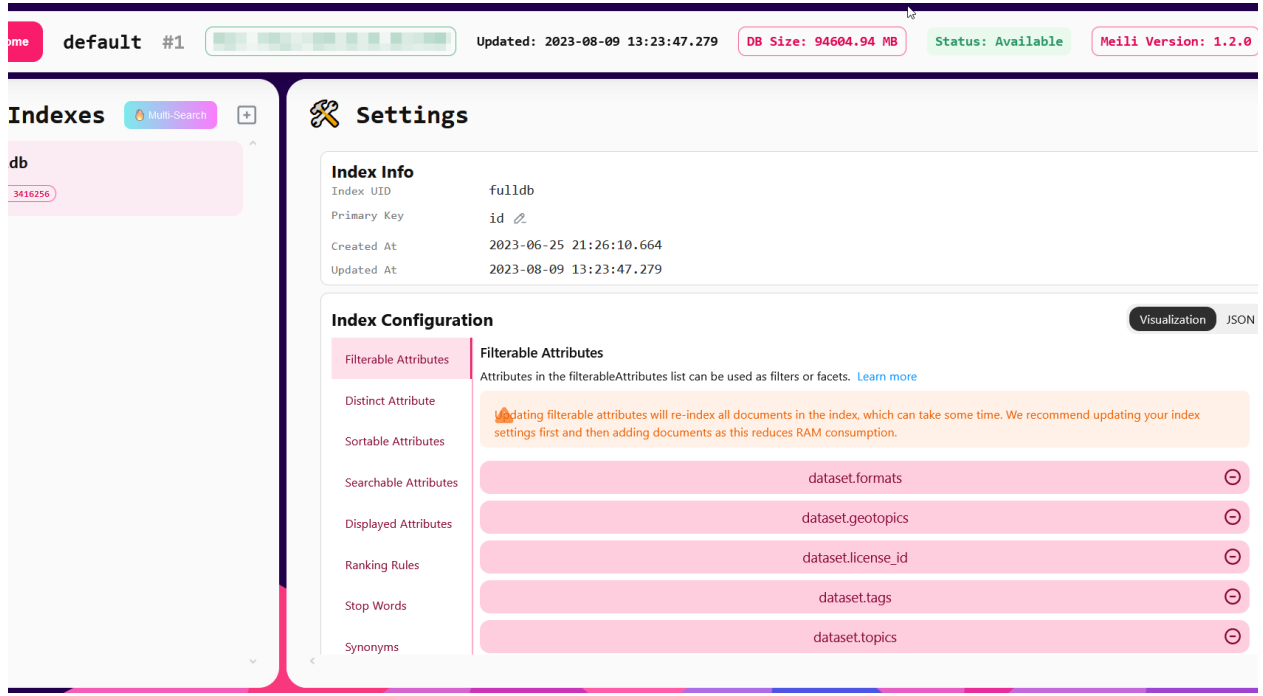
Требует анализа непосредственно файла/API с данными. Высокая вариантивность, специальная онтология

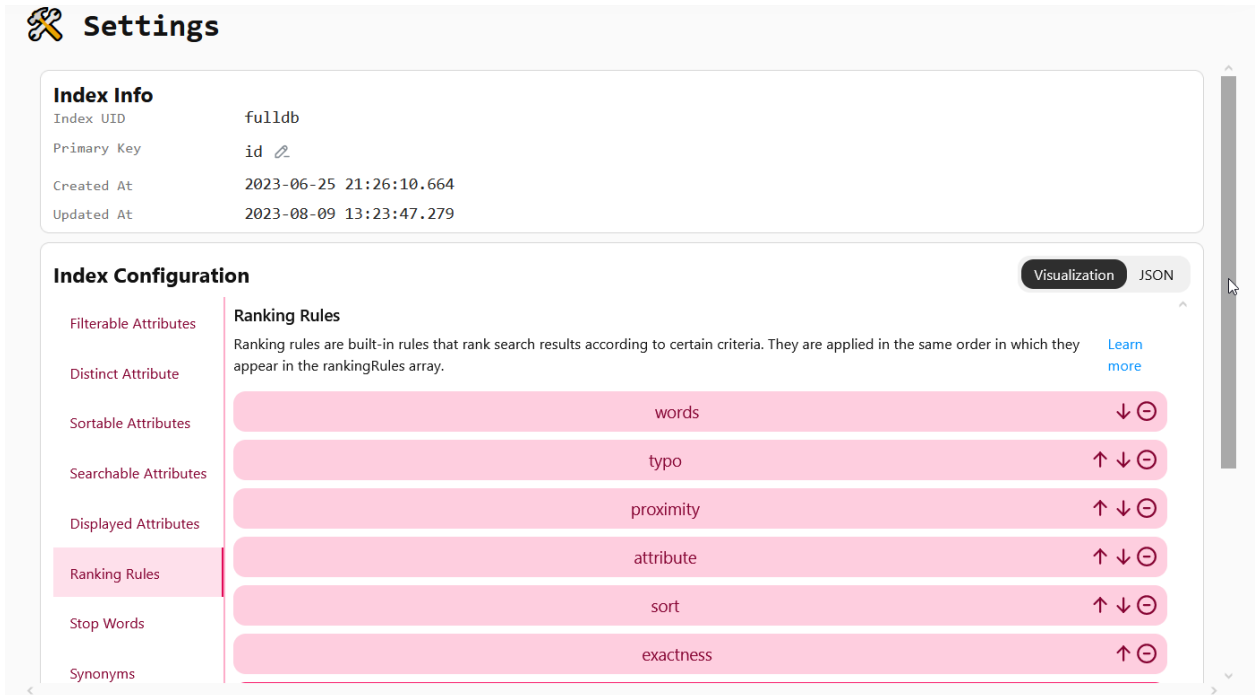
Этап 4. Поиск

Поиск

- Изначально предполагался Elastic/OpenSearch
- База поискового индекс в MongoDB, поисковый движок Meilisearch (пока на старой версии 1.2)
- Плюсы: прекрасная скорость, поддержка фасетов из коробки, большое число языков, отличная техподдержка
- Минусы: долгое индексирование, огромный индекс без сжатия, высокие требования к RAM
- 2 часа на индексирование 3.3 миллионов наборов данных, база в 5.3 ГБ, 94.6 ГБ индекс (коэффициент 17.8)







Админка Meilisearch

Действительно ли Meilisearch наиболее подходящий движок? Какие есть альтернативы?

Нерешённое

Превышение объёмов данных

При росте базы до 50 миллионов наборов данных ожидаемый размер поискового индекса Meilisearch достигнет 1.5 терабайта. Неизвестно насколько Meilisearch потянет такой объём

Недостаточное качество поиска

Сейчас поиск основан на данных из первоисточников и работы алгоритмов распознавания ряда атрибутов.

Недостаточная релевантность

Когда нет PageRank релевантность может не сохраняться, особенно с ростом числа наборов данных.

Результат на сегодня

Получилось

- Создать крупнейший реестр каталогов данных в мире
- Собрать метаданные из более чем 1500 источников
- Сделать альфа версию поискового индекса
- Сделать демо поиска

Пока не получилось

- Сделать глубокий поиск по типам данных и семантическим типам данных
- Достигнуть отметки в 50 миллионов наборов данных

Неожиданное

- Очень быстрый поиск на базе Meilisearch

http://demo.commondata.io

Вопросы?

email: ivan@begtin.tech
Telegram: https://t.me/begtin