

Каталог и озеро данных на базе MongoDB

Собираем технологический стек по кусочкам и работаем с неуправляемыми источниками данных

Кто я и почему об этом рассказываю?

1. Много лет
проектировал
корпоративные и
государственные
data-rich проекты



2. Возглавляю НКО
Инфокультура и
Ассоциацию
участников рынка
данных. Продвигаю
открытые данные в
России

3. Люблю
программировать, но
времени мало поэтому
пишу “ленивый код”

4. Развиваю
технологический
стартап
APICrafter/DataCrafter

Каталоги данных

Какие каталоги данных бывают

Каталоги данных

Порталы
открытых
данных

Научные
репозитории

Корпоративные
каталоги

Отраслевые
каталоги

Data.gov

Hubofdata.ru

Harvard
DataVerse

Управление
НСИ

nsi.rosminzdrav.ru

Каталоги данных
для машинного
обучения

kaggle.com/datasets

ЕСИМО

Ситуация с классическими каталогами [мета]данных

- Узкозаточены под конкретные задачи
- Не умеют (или очень дорого умеют) идентифицировать типы полей
- Плохо поддерживают не-табличные данные
- Не умеют или умеют дорого автоматически анализировать качество данных

Современные публичные каталоги данных



Data.world

- Большой акцент на визуализации
- Удобное рабочее место
- Пытаются стать Github'ом для данных
- Формируют сообщество



doltHub

- Git-like база данных
- Пытаются стать Github'ом для данных
- Формируют сообщества



Bit.IO

- Делитесь не датасетами, базами данных (с)
- Позволяют сразу создавать свои онлайн СУБД и делиться ими с другими

Какой каталог хотелось бы сделать

- 1 Универсальный публичный каталог данных
- 2 Для коммерческого и некоммерческого использования
- 3 Автоматизировать всё что можно автоматизировать
- 4 Научиться работать с данными любого качества
- 5 Собрать много, очень много, данных
- 6 И, конечно, запихнуть туда кучу наработок по работе с данными

**Какие проблемы
пришлось решать**

Неуправляемые источники данных

- На источники данных сложно или невозможно повлиять.
- Данные публикуются в самых разных форматах от JSON до DBF, XML, Excel, Word и тд.
- Часто данные можно собирать только парсерами HTML или PDF или иных документов
- Даже для открытых данных владельцы могут предусматривать ограничения на их выгрузку

```

▼ placer:
  ► fullName: "ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕН_РСТВЕННЫЙ УНИВЕРСИТЕТ\"
    iko: "87801002274780101001"
    inn: "7801002274"
    kpp: "780101001"
  ► legalAddress: "199034, г Санкт-Петербур_ниверситетская, дом 7/9"
    ogrn: "1037800006089"
    okato: "40263561000"
    okopf: "75103"
    okopfName: "Федеральные государственные бюджетные учреждения"
    okpo: "02068516"
    phone: "7(812)3200710"
    postalAddress: "199034, Санкт-Петербург, Университетская, дом 7/9"
  ► shortName: "САНКТ-ПЕТЕРБУРГСКИЙ ГОСУ_Й УНИВЕРСИТЕТ ИЛИ СПБГУ"
▼ purchaseTypeInfo:
  code: "30000"
  ► name: "Закупка у единственного _одрядчика, исполнителя)"
▼ contractPositions:
  ▼ 0:
    ► okei: {}
    ► okpd2: {}
    guid: "2870d645-c004-4786-bf40-c2fdbae944e6"
    ► name: "Расшифровка рукописей и _олами). Объем - 48 стр."
    ordinalNumber: 1
    typeObjectPurchase: "S"
    impossibleToDetermineAttr: false
    qty: 1
  ▼ 1:
    ► okei: {}
    ► okpd2: {}
    guid: "3193fd7e-20e2-43b1-afed-62d5ad4143fb"
    ► name: "Расшифровка рукописей и _ годы). Объем - 52 стр."
    ordinalNumber: 2
    typeObjectPurchase: "S"
    impossibleToDetermineAttr: false

```

● Данные не плоские!

Почти все источники данных содержат вложенные объекты и подобъекты и закодированы в JSON/XML

● Нет готового ПО

Корпоративные каталоги не заточены под неплоские данные и ПО для интеграции данных и их сбора и преобразования тоже почти всё про таблицы

Документировать данные - это дорого!

- В среднем 48 полей на один набор данных
- Только открытых данных на госпорталах не меньше 50 тысяч, а скорее 100 тысяч
- Это до **5 миллионов полей!**

Эксперименты

Автоматическая идентификация дат

- **QDDate** (quick dirty dates) <https://github.com/ivbeg/qddate>
- Очень быстро распознает даты в виде текста в более чем 300 способов написания
- Поддерживает даты написанное на русском, немецком, испанском и других языках
- Проверялось на десятках тысяч сайтов для извлечения новостей
- Написана без регулярных выражений, всё распознается с помощью PyParsing и “упихивания” шаблонов во вручную запрограммированный индекс их перебора (всё не по канонам, но работает)

Автоматическое извлечение новостей

- **Newsworker** <https://github.com/ivbeg/newsworker>
- Очень полезна для работы с сайтами всяких международных структур и госсайтами
- Позволяет извлекать новости из сайтов не поддерживающих RSS
- Из этой библиотеки была выделена QDDate
- Проверялось на десятках тысяч сайтов для извлечения новостей

Автоматическое создание API

- **APIReady** <https://github.com/ivbeg/apiready>
- Изначально решалась задача автогенерации API на основе любых в виде датасета (csv/json/xml) и тд. Так чтобы сразу API из коробки без доп. усилий
- Автоматически анализирует JSON или CSV файл и преобразует в базу MongoDB
- Находит колонки/поля являющиеся словарями и выделяет их в отдельные коллекции
- Делает данные доступными через API

Преобразование и обработка не-плоских данных

- **Undatum** <https://github.com/datacoon/undatum/>
- Попытка создать “швейцарский нож” для работы с JSON lines и BSON
- Похож на csvkit, с оговоркой что данные не плоские
- Автоматически анализирует, создаёт схемы данных
- Включает код автопроверки идентификаторов (email, INN, OGRN и др.)

Автоматическая классификация и документирование

- **Mongo2Md**
<https://github.com/datacoon/mongo2md>
- Автоматически создает шаблоны для описания полей коллекций в MongoDB
- Позволяет вручную заполнить шаблоны и на их основе сгенерировать документацию в Markdown

- 1Запускаем “mongo2md prepare -o projpath dbname collname”
- 2Получаем файлы: tables.csv, %collname%_fields.csv, %collname%_schema.json, %collname%_example.json
- 3Редактируем tables.csv и %collname%_fields.csv, добавляем туда описания в поля description
- 4Запускаем “mongo2md document -o projpath”. Смотрим файлы projpath/markdown/tables.md

```
me;type;description;sample;class
d;string;Уникальный суррогатный ключ;607f334cec3f5df685ff23ae;
me;string;;Государственное бюджетное учреждение здравоохранения «Волгоградский областной медицинский информационно-аналитический центр»;orgname
ail;string;Адрес электронной почты;ca@vomiac.ru;email
ort_name;string;;ГКУЗ «ВОМИАЦ»;orgname
fo_url;string;;http://vomiac.ru/ca.html;url
rtlist_url;string;;;url
dress;dict;Адрес;;
dress.country;string;;RU;
dress.region;dict;;;
dress.region.code;string;;34;
dress.region.name;string;;Волгоградская область;rusregion
dress.index;string;;400005;postindex
dress.street_house;string;;Проспект им В.И.Ленина 56a;
dress.city;string;;Волгоград;
ogram_complexes;string;;{'complex': [{'preudonim': 'ГБУЗ ""ВОМИАЦ""', 'ep_class': 'KC2', 'address': {'country': 'RU', 'region': {'code': '34
n;string;Код ИНН (Идентификационный номер налогоплательщика);3443902211;inn
rn;string;Код ОГРН (Общероссийский государственный регистрационный номер);1023402981777;orgname
gnum;string;;308;
cred_status;dict;;;
cred_status.status;string;;Действует;
cred_status.valid_since;datetime;;2013-09-24 00:00:00;date
cred_status_history;dict;;;
cred_status_history.accred_status;list of [array];;
cred_status_history.accred_status.status;string;;Действует;
cred_status_history.accred_status.valid_since;datetime;;2013-09-24 00:00:00;date
```

```
# Data structures
## Collection accredituс, Реестр аккредитованных удостоверяющих центров
Реестр аккредитованных удостоверяющих центров с портала УЦ Минсвязи России

Name / Type / Description /
/ -----:| -----:| -----:|
/ _id / string / Уникальный суррогатный ключ
/ accred_status / dict as [accred_status](accred_status) /
/ accred_status_history / dict as [accred_status_history](accred_status_history) /
/ address / dict as [address](address) / Адрес
/ certlist_url / string /
/ email / string / Адрес электронной почты
/ info_url / string /
/ inn / string / Код ИНН (Идентификационный номер налогоплательщика)
/ name / string /
/ ogrn / string / Код ОГРН (Общероссийский государственный регистрационный номер)
/ program_complexes / string /
/ regnum / string /
/ short_name / string /

### Example
'''javascript
{
  "_id": {
    "$oid": "5fe438956279004de7ed0edd"
  },
  "accred_status": {
    "status": "Действует",
    "valid_since": {
      "$date": "2013-09-24T00:00:00.000Z"
    }
  },
  "accred_status_history": {
    "accred_status": {
      "status": "Действует",
      "valid_since": {
        "$date": "2013-09-24T00:00:00.000Z"
      }
    }
  }
}
```

Data stuctures

Collection accredituс, Реестр аккредитованных удостоверяющих центров

Реестр аккредитованных удостоверяющих центров с портала УЦ Минсвязи России

Name	Type	Descriptor
_id	string	Уникальный суррогатный ключ
accred_status	dict as accred_status	
accred_status_history	dict as accred_status_history	
address	dict as address	Адрес
certlist_url	string	
email	string	Адрес электронной почты
info_url	string	
inn	string	Код ИНН (Идентификационный номер налогоплательщика)
name	string	
ogrn	string	Код ОГРН (Общероссийский государственный регистрационный номер)

Автоархивация данных из API

- **APIBackuper** <https://github.com/ruarxive/apibackuper>
- Позволяет извлекать данные из API эндпоинтов перебирая их по определённым правилам
- Заодно умеет архивировать файлы, на которые в данных есть ссылки
- Отдаёт данные в JSON lines
- Изначально создавалось для архивации данных из портала budget.gov.ru (там всё в API, нормальные данные не скачать)

- 1 Готовим конфиг `aribackuper.cfg`
- 2 Запускаем `"aribackuper estimate"` для проверки и оценки сроков
- 3 Запускаем `"aribackuper run"` для выкачки данных
- 4 Запускаем `"aribackuper export data.jsonl"` чтобы получить данные в JSON lines

```
5 [project]
6 description = Moscow Government official documents
7 url = https://www.mos.ru/api/documents/v2/frontend/json/ru/documents
8 http_mode = GET
9 work_modes = full,incremental,update
10 iterate_by = page
11
12 [params]
13 page_size_param = per-page
14 page_size_limit = 50
15 page_number_param = page
16
17 [data]
18 total_number_key = _meta.totalCount
19 data_key = items
20 item_key = id
21 change_key = updated_at
```

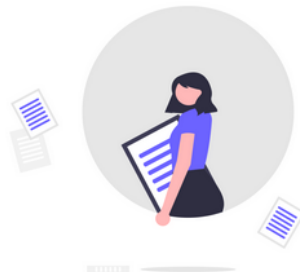
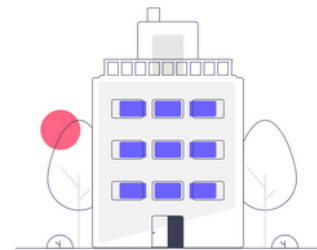
```
9 {"id": 243457220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
10 {"id": 243458220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
11 {"id": 243459220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
12 {"id": 243460220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
13 {"id": 243461220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
14 {"id": 243462220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
15 {"id": 243463220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
16 {"id": 243464220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
17 {"id": 243465220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
18 {"id": 243466220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
19 {"id": 243467220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
20 {"id": 243468220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
21 {"id": 243469220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
22 {"id": 243470220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
23 {"id": 243471220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
24 {"id": 243472220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
25 {"id": 243473220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
26 {"id": 243474220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
27 {"id": 243475220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
28 {"id": 243476220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
29 {"id": 243477220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
30 {"id": 243478220, "title": "Оповещение о начале общественных обсуждений по внесению изменений в правила землепользовани
31 {"id": 243479220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
32 {"id": 243480220, "title": "Оповещение о начале общественных обсуждений по внесению изменений в правила землепользовани
33 {"id": 243481220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
34 {"id": 243482220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
35 {"id": 243483220, "title": "Оповещение о начале общественных обсуждений по проекту внесения изменений в правила землепо
```

Что получилось

Наборы данных по темам

**Госфинансы**
116 наборов данных

Данные о государственных финансах, бюджетах, субсидиях, госконтрактах и иных бюджетных отношениях.

**Документы**
6 наборов данных
Документы, включая их тексты и ссылки на первичные опубликованные файлы**Здания**
8 наборов данных
Сведения о зданиях, земле, объектах кадастрового учета**Интернет и сетевая инфраструктура**
2 наборов данных
IP адреса, сведения WHOIS, домены, инфраструктура связи и иное с этим**Информационная безопасность**
4 наборов данных
Удостоверяющие центры, защита информации, защита данных и тд.**Информационные технологии**
6 наборов данных
Всё что связано с информационными технологиями, ИТ компаниями, ИТ продуктами

Что делает DataCrafter?



Сбор данных

Сейчас доступны более чем 200 крупных источников (базы государственных и муниципальных закупок и контрактов, ЕМИСС, Центральный банк, ФНС России, реестры среднего и малого бизнеса, Рособrnнадзора и десятки других)



Преобразования в удобные форматы

Накоплен большой опыт по очистке и верификации данных, в том числе в автоматизированном режиме



Актуализация

Регулярность обновления данных в зависимости от набора данных – от 1 дня до 1 месяца. Например, данные о госзакупках обновляются ежедневно



Единая среда

Доступ ко всем данным предоставляется через удобные API и слепки данных



Внешний портал

Возможность поработать с данными до заключения договора на портале **data.apicrafter.ru**

Ценностное предложение

API

Доступ через стандартизованное REST API с поиском записей и получение списка записей по запросу

Сборки данных

Выгрузка полной копии данных в виде базы данных для MongoDB или резервной копии данных в формате JSON lines

Доступ к СУБД

Прямой доступ к базам данных при развертывании на выделенном сервере или в инфраструктуре заказчика

Интерфейс под заказ

Разработка интерфейсов выгрузки в форматах CSV/XML/JSON или веб интерфейсов под задачи заказчика

Примеры данных

- Государственные и муниципальные контракты с 2007 года
- Базы официально опубликованных нормативно-правовых документов
- Государственные и муниципальные закупки с 2013 года
- Протоколы по государственным и муниципальным закупкам с 2013 года
- Закупки малого объёма по субъектам федерации
- Документы стратегического планирования с портала ГАС Управление
- Реестры поставщиков и заказчиков
- ЕГРЮЛ
- Статрегистр юр. лиц
- Бухгалтерские балансы юр. лиц
- Бухгалтерские балансы госучреждений
- Индикаторы статистики Росстата (ЕМИСС)
- Ведомственная статистика ФНС, Минкультуры, Генпрокуратуры и тд.

Объёмы данных (Data Crafter)

- **533 коллекции в 391 пакетах данных**

По организациям, медицине, госфинансам, финансам, медицине и тд.

- **246 ГБ структурированных данных**

Данные подготовлены для доступа через API и сборок данных.

Объёмы данных (специализированные базы)

- **Крупные базы данных**

ЕГРЮЛ, база госзакупок и госконтрактов, база сведений о госучреждениях

- **Около 1 ТБ структурированных данных**

Доступны через специализированные API

- **Проходят специальную подготовку**

Оптимизация под производительность, загрузка в PostgreSQL и разработка специальных API

Объёмы данных (в ожидании к загрузке)

- **200+ каталогов данных и более 100 тысяч наборов данных**

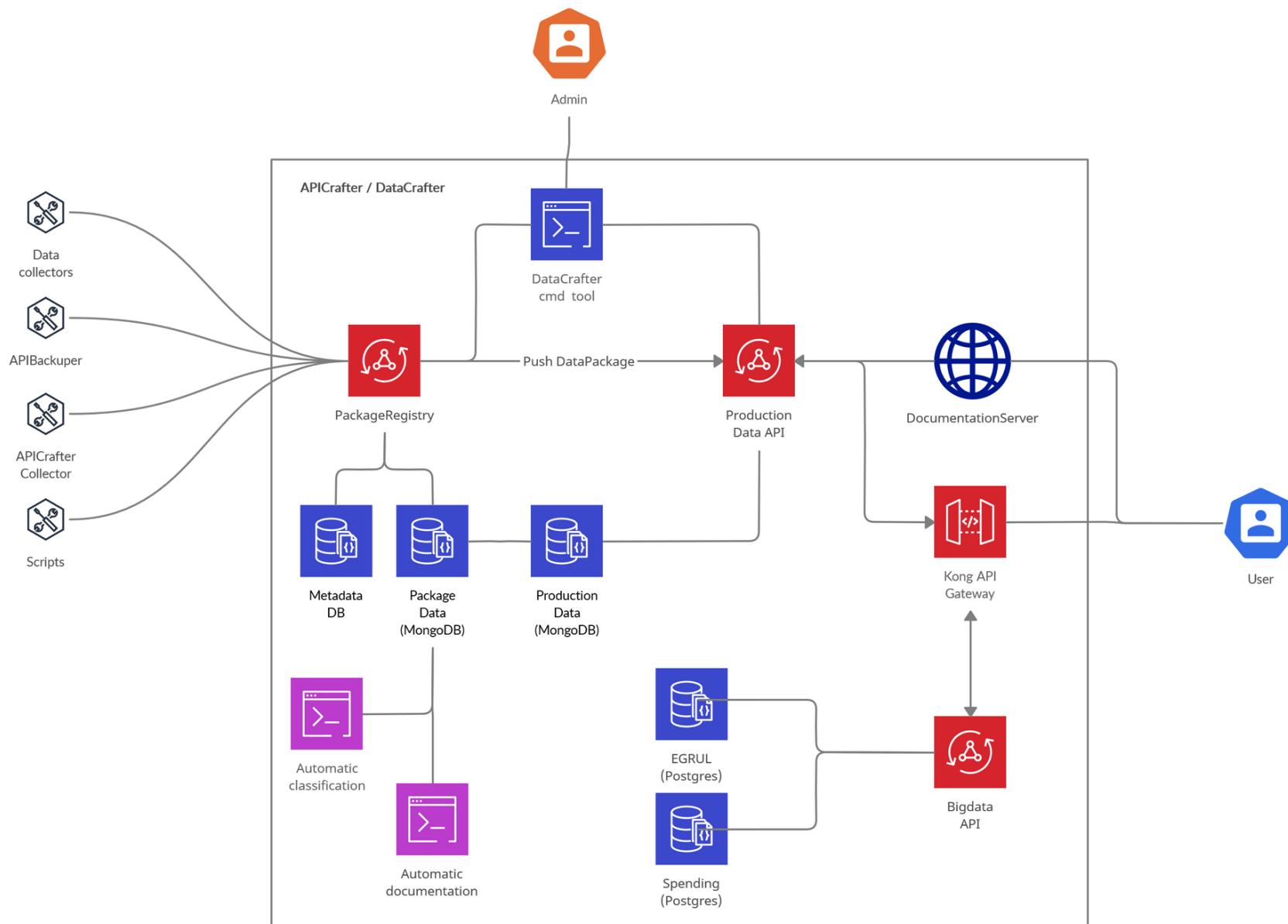
Автоматическая загрузка данных и крупнейших каталогов

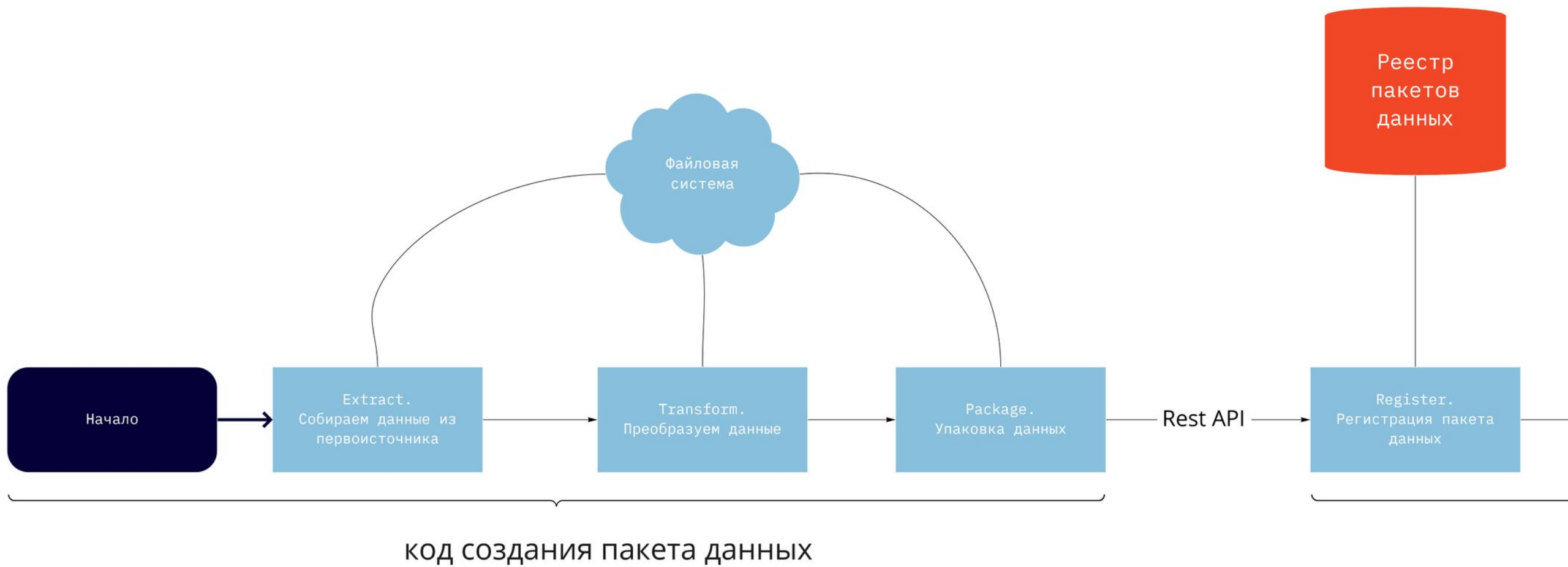
- **Около 5 ТБ структурированных данных**

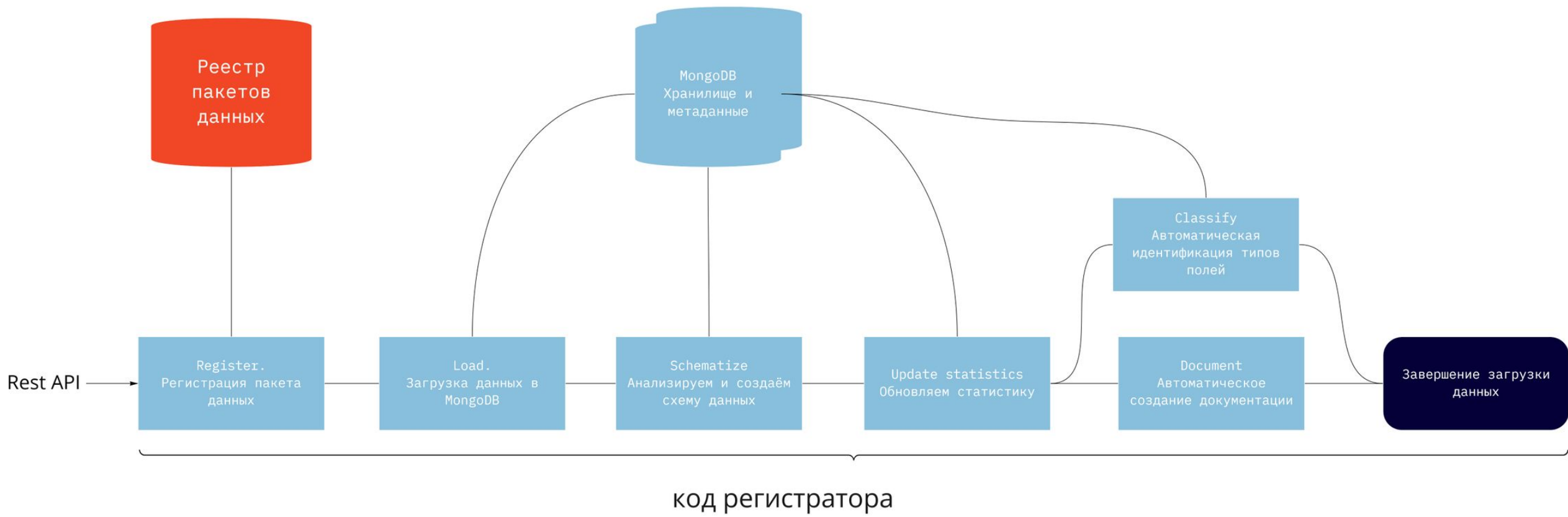
Базы документов, адресные реестры и иные материалы

- **Проходят оценку востребованности**

Не все из этих данных реально нужны в регулярной, повседневной работе.







Технологический стек

- Хранилище: MongoDB + Postgres
- Управление API: Python Eve + Kong
- Доступ к API: OpenAPI/Swagger
- Экспорт данных: BSON + JSON lines
- Фронтэнд: Python+Flask
- Сбор данных: своя разработка
- Data pipelines: своя разработка
- Классификация данных: своя разработка

data{crafter}

MongoDB vs. все остальные

SQL

Четкая структура данных
Требует проектирования
структуры данных под плоские
таблицы

NoSQL (MongoDB)

Наиболее популярная NoSQL база данных с большим сообществом,
библиотеками и инструментами

Почему MongoDB?

- Неуправляемые первоисточники данных
- Не-плоские данные по умолчанию
- Меньше издержек на ручное поддержание схем данных
- Многолетний опыт работы*

* больше 20 проектов на MongoDB ранее сделано с данными до 500 Гб в транзакционной базе

Пакеты данных как стратегия

- Есть много стандартов: Frictionless Data, BagIt, DataCrate, Research Objects, Metapack, ReproZIP, DSPL и др.
- Почти все работают только с плоскими данными/таблицами
- Ориентированы на исследователей
- Самый зрелый - [FrictionlessData.io\](https://frictionlessdata.io/)
- Ограничения: не реалтайм, не big data (не терабайты), инкрементальные обновления ограничены

Пакеты данных

- пришлось разработать свой под не-плоские данные и иную специфику
- стандартизирует описание наборов и баз данных
- позволяют получать данные откуда угодно

```
{
  "name": "resourceprj",
  "title": "Выплаты нефтегазовыми и иными ресурсодобывающими компаниями правительствам стран",
  "description": "Данные из проекта Resource Projects со сведениями о выплатах нефтегазовыми и иными",
  "licenses": [
    {
      "name": "ccby40",
      "path": "",
      "title": "Creative Commons"
    }
  ],
  "sources": [
    {
      "name": "resproj",
      "title": "Resource Projects",
      "path": "https://resourceprojects.org"
    }
  ],
  "keywords": [
    "organizations"
  ],
  "version": "20210716-10-00-00",
  "package_type": "full",
  "topics": [
    {
      "id": "organizations",
      "name": "Организации"
    }
  ],
  "spec_version": "1.0",
  "spec_type": "crafter-datapackage",
  "documentation": [],
  "collections": [
    {
      "name": "companies",
      "title": "Ресурсные компании",
      "description": "Ресурсные компании",
      "dbname": "resprojects",
      "collname": "companies",
      "path": "data/companies.jsonl",
      "compression": null,
      "format": "jsonl",
      "media": "application/json"
    }
  ]
}
```

Классификация полей данных

В цифрах

- 69 типов полей
- 23,5% идентифицируется автоматически

Как работает?

Несколько сотен правил
Ручная перепроверка
Анализ по содержанию полей и их названию

Где посмотреть?

<https://data.apicrafter.ru/class>

Вызовы

Если загрузить все наборы данных, то будут миллионы полей очень разного типа

Что дальше?

Алгоритмы автоматического выявления неочевидных, повторяющихся типов данных

Классы данных

Код	Название	Всего полей
address	Адрес	193
bankaccount	Банковский счёт (расчётный, депозитный, корреспондентский и тд.)	5
bik	Банковский идентификационный код	7
boolean	Логический тип данных (boolean)	495
budgetcode	Код бюджета Российской Федерации	23
budgetname	Наименование бюджета Российской Федерации	64
date	Дата	3
datetime	Дата и время	1036
email	Email	65
fedgrbs	Главные распорядители бюджетных средств (федеральный бюджет)	14
geopoint	Географическая точка	6
guid	Уникальный идентификатор GUID	264
iku	Идентификационный код заказчика	2
inn	Идентификационный номер налогоплательщика	199
ipaddr	IP адрес (может быть IPv4 или IPv6)	2
ipsubnet	Подсеть IP адресов	1
ipv4	Адрес IPv4	1
ipv4subnet	Подсеть IP адресов IPv4	1
ipv6	Адрес IPv6	1
kbk	Код бюджетной классификации	3
kladr	Код КЛАДР	5

Общероссийский классификатор органов государственного управления

Поля

Название поля	Тип поля	Таблица	Набор данных
details.okogu_code	string	Поставщики зарегистрированные в агрегаторе закупок Березка	Поставщики зарегистрированные в агрегаторе закупок Березка
info.okoguCode	string	Реестр участников и неучастников бюджетного процесса	Реестр участников и неучастников бюджетного процесса
protection_organization.okogu	string	Диссертация	ЕГИСУ НИОКТР (rosrid.ru). Обновлённая система
author_organization.okogu	string	Диссертация	ЕГИСУ НИОКТР (rosrid.ru). Обновлённая система
customer.okogu	string	Результаты интеллектуальной деятельности	ЕГИСУ НИОКТР (rosrid.ru). Обновлённая система
executors.okogu	string	Результаты интеллектуальной деятельности	ЕГИСУ НИОКТР (rosrid.ru). Обновлённая система
customer.okogu	string	Научно-исследовательские, опытно-конструкторские, технологические работы	ЕГИСУ НИОКТР (rosrid.ru). Обновлённая система
executor.okogu	string	Научно-исследовательские, опытно-конструкторские, технологические работы	ЕГИСУ НИОКТР (rosrid.ru). Обновлённая система
company.okogu	string	Государственные заказчики системы госзаказа города Москвы	Государственные заказчики системы госзаказа города Москвы
requestInfoDomain.authApprove.okogu	string	Документы стратегического планирования	Документы стратегического планирования
requestInfoDomain.authImplement.okogu	string	Документы стратегического планирования	Документы стратегического планирования
code	string	Общероссийский классификатор органов государственной власти и управления	Общероссийский классификатор органов государственной власти и управления
customerOkogu	string	Результаты интеллектуальной деятельности	ЕГИСУ НИОКТР. Система до 2020 года
details.okogu_code	string	Заказчики зарегистрированные в агрегаторе закупок Березка	Заказчики зарегистрированные в агрегаторе закупок Березка
okogu	dict	Организации НКО	База некоммерческих организаций

Архитектурные решения и их последствия

- **MongoDB для хранения данных и метаданных**

- Для каждой таблицы отдельная коллекция. Что делать когда будет 100500+ таблиц?
- Возможно стоит построить свой schema-less движок вокруг другого хранилища?

- **Python как основной язык разработки**

- Быстрая разработка, но не самая быстрая обработка данных
- Решается отделением пакетов данных

- **Python Eve для автоматического создания API**

- Не поддерживает полнотекстовый поиск и сложные запросы
- Не так просто расширяется
- На самом деле не schema-less, надо генерировать схемы налету

- **Классификация данных**

- Надо автоматизировать контроль качества идентификации типов данных и автодокументирования

Непройденные развилки

1

Git для данных

Идти ли по пути Dolthub'a и внедрять git для данных, с коммитами изменений?

2

Данные пользователей и пользовательский интерфейс

Дать возможность пользователям грузить данные или пойти по пути только дата продуктов и сервисов?

3

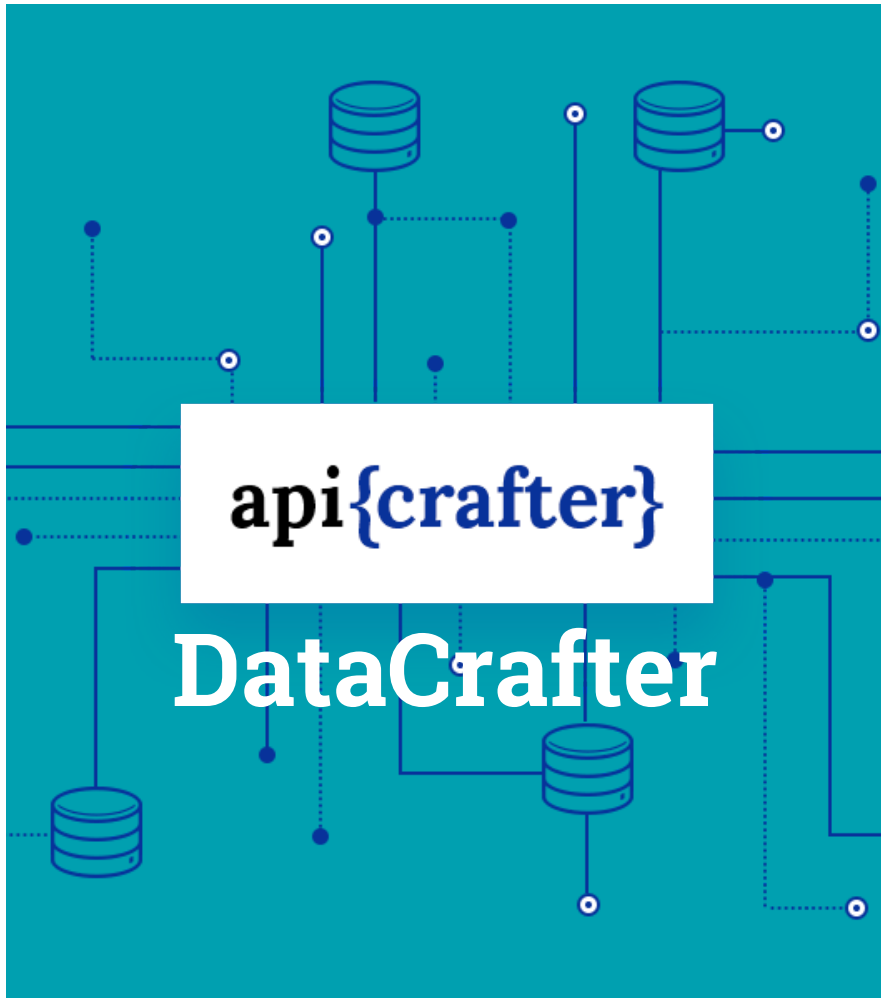
ELT и Cloud-first

Переходить ли к ELT или исходить из того что при работе через пакеты данных Extract и Transformation на не волнуют?

Планы (много планов)

- Загрузить 100500+ наборов данных
- Закончить и сделать публичным автоклассификатор и автодокументатор данных
- Добавить интеграционные интерфейсы (PowerBI, Excel, Google Sheets, Tableau)
- Больше клиентов, больше пользователей!

Контакты



По всем вопросам обращайтесь:

apicrafter@apicrafter.ru

Иван Бегтин

Запланированные данные

Электронный бюджет

Около 150 пакетов данных, около 150 ГБ.

Данные по госрасходам, госдоходам, участникам бюджетного процесса и так далее.

Порталы открытых данных

Около 15000 наборов данных, около 1000 ГБ.

Данные из региональных порталов открытых данных

Реестры организаций

Сотни реестров организаций таких как торговые реестры, реестры поставщиков, реестры лицензий

Около 500 наборов данных, около 100 ГБ

Особенно значимые данные

Госкаталог культурного фонда, ФИАС, ведомственная статистика, муниципальная статистика

Общий объём около 500 ГБ, около 10 баз данных

Базы нормативных документов

Базы данных НПА федерального уровня и субъектов федерации

Более 1 миллиона документов, более 3 терабайт данных