Информационная культура

Open source for Opendata

Методические рекомендации по использованию решений с открытым кодом при работе с открытыми данными

Иван Бегтин, Ксения Орлова,

АНО «Информационная культура»

Оглавление

Введение

Почему мы создали эти рекомендации и кому они полезны

Движения за открытый код и открытые данные имеют множество исторических пересечений. Многие из активистов за открытость данных пришли в эту деятельность из других движений за открытость, таких как открытый код, открытое оборудование и открытые знания. В то же время, не все участники сообществ по открытости данных знают о существовании инструментов с открытым кодом, которые могут помочь им в ежедневной работе. Эти рекомендации собраны для того, чтобы описать способы использования таких инструментов и помочь специалистам разного профиля и уровня работы с данными.

Для органов власти и государственных учреждений

Органы власти и государственные учреждения одновременно являются и владельцами, и потребителями данных, зачастую не обладая необходимыми финансовыми ресурсами для использования коммерческих продуктов. Открытый код позволяет существенно снизить издержки при запуске любой инициативы, связанной с открытыми данными, поэтому инструменты в этих рекомендациях отобраны с фокусом на те задачи, которые возникают в процессе работы с данными у государственных структур.

Для пользователей данных

В рекомендациях собраны примеры инструментов, позволяющих довольно быстро настроить собственные процессы обработки, анализа и визуализации данных. Кроме того, рекомендации для владельцев данных помогут настроить диалог с пользователями данных, что поможет создать эффективную инфраструктуру данных и снизить издержки, например, при подготовке данных к публикации или разработке порталов данных.

Виды деятельности при работе с данными

Несмотря на то, что многие инструменты универсальны и используются для многих задач, есть несколько видов деятельности работы с данными, для которых они применяются в первую очередь. Можно сказать, что это основная специализация именно этого инструмента/продукта.

Мы определяли эти виды деятельности исходя из собственного опыта и с оглядкой на классификацию инструментов в таких сервисах, как Stackshare.io1 и его аналогах. Формируя текущий список инструментов, мы выделили следующие категории:

1 Stackshare: Track and collaborate on tech stack decisions - stackshare.io

- подготовка и упаковка данных;
- создание порталов и каталогов данных;
- сбор и извлечение данных;
- очистка и контроль качества данных;
- обработка данных;
- аналитика и Business Intelligence;
- работа с геоданными;
- машинное обучение и искусственный интеллект;
- визуализация данных.

А также в карточке каждого инструмента указываются дополнительные категории, к которым он может быть отнесён. Этот перечень категорий не является строго зафиксированным и окончательным и будет меняться в следующих версиях этого руководства.

Как выбрать нужные инструменты

При работе с данными выбор инструмента, в первую очередь, определяется теми задачами, которые предполагается выполнять. Мы сгруппировали инструменты по видам деятельности и подготовили рекомендации по каждому из них, чтобы этот выбор упростить.

Другой важный критерий выбора инструментов — язык программирования. Поскольку чаще всего инструмент будет использоваться в программной среде, и важно, чтобы его язык разработки соответствовал рабочему окружению разработки (IDE) вашей команды.

К программной среде также относится лицензия, под которой публикуется исходный код инструмента. Если вы хотите интегрировать его в ваш продукт, то важно знать имеете ли вы такую возможность, и какие ограничения это наложит на ваш исходный код: должны ли вы будете его опубликовать, например.

И, безусловно, значение имеет визуальная среда работы инструмента. Некоторые инструменты могут использоваться как веб-приложения и вам потребуется установить их на отдельный сервер, другие инструменты можно установить как настольные приложения, и во многих случаях это будут инструменты для командной строки (CLI) — работа с ними, как правило, требует соответствующего опыта в оболочках Unix или Terminal для Windows.

Виды деятельности при работе с данными

Подготовка и упаковка данных

Подготовка и упаковка данных — важные процессы при планировании публикации данных, включающие сбор метаданных о каждом наборе данных и выбор форматов и стандартов публикации данных. Использование специальных инструментов позволяет публиковать данные в соответствии с формализованными описаниями. Подготовка данных особенно большое значение имеет в тех случаях, когда данные нуждаются в регулярном обновлении, например, для справочников и иных референсных реестров.

Для кого это важно: Владельцы данных

Сложность использования: Продвинутая

Рекомендации: DataCurator не требует технических навыков и Frictionless Data для задач, требующих автоматизации подготовки данных.

Data Curator

Назначение инструмента	Подготовка и упаковка данных
Уровень сложности	Простой
Доступен исходный код	Да
Тип продукта	Настольное приложение
Ссылка на сайт	github.com/qcif/data-curator
Ссылка на открытый код	github.com/qcif/data-curator
Лицензия на код	MIT License
Язык разработки	Python
Поддерживаемые типы данных	Любые типы данных

Data Curator — это простой настольный редактор данных, помогающий в описании и валидации данных, а также в их дальнейшем распространении. Он использует стандарт Frictionless Data в части описания схемы таблиц (Table Schema) и пакетов данных (Data Package) для импорта и экспорта данных. Разработан в ODI Queensland (Австралия).

Лучшие практики использования: Data Curator используется для ручной подготовки данных. Он упаковывает данные в формате Data Packages для публикации с заполнением метаданных и имеет мало альтернатив.

Frictionless Data

Назначение инструмента	Подготовка и упаковка данных
Уровень сложности	Сложный
Доступен исходный код	Да
Тип продукта	Командная строка
Ссылка на сайт	frictionlessdata.io
Ссылка на открытый код	github.com/frictionlessdata/project/
Лицензия на код	MIT License

Язык разработки	Python, R, Javascript, Swift, PHP, Java, Clojure, Julia, Go, Ruby, Javascript
Поддерживаемые типы данных	CSV, JSON, Data Package

Frictionless Data — это одновременно и стандарт публикации табличных открытых данных, и широкий набор открытых инструментов по подготовке и упаковке данных. Используется в десятках порталов открытых данных по всему миру, включая портал открытых данных Правительства Великобритании (data.gov.uk). Создан и развивается командой Open Knowledge Foundation.

Лучшие практики использования: Frictionless Data — это реализация одноименного стандарта управления метаданными для наборов данных. Похож на лучшие практики в области упаковки цифрового контента, в частности — стандарт и спецификацию BagIt2. Frictionless Data не имеет других открытых альтернатив для работы с табличными данными общего назначения.

Создание порталов и каталогов открытых данных

Порталы открытых данных необходимы потребителям данных, чтобы нужные им наборы данных можно было найти наиболее удобным образом, узнать все необходимые сведения, иметь возможность получить данные как вручную, так и автоматически. Порталы создаются владельцами данных, обладающими достаточно большим числом наборов данных, начиная с нескольких десятков наборов данных. Порталы и каталоги открытых данных сочетают множество функций: предпросмотр данных, визуализация, контроль метаданных качества данных. Это упрощает взаимодействие внешних пользователей с порталами и каталогами открытых данных.

Для кого это важно: Владельцы данных

Сложность использования: Продвинутая

2 The BagIt File Packaging Format (V1.0) tools.ietf.org/html/rfc8493

Рекомендации: CKAN обладает наиболее продвинутыми возможностями по разработке порталов открытых данных — если есть технические возможности его развернуть, то это лучший выбор. DKAN выступает в роли более простой альтернативы, удобной в условиях дефицита технических ресурсов.

CKAN

Назначение инструмента	Каталоги и порталы данных
Уровень сложности	Сложный
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	ckan.org
Ссылка на открытый код	github.com/ckan/ckan
Лицензия на код	GNU Affero General Public
Язык разработки	Python
Поддерживаемые типы данных	Любые типы данных

CKAN — это программный продукт для разработки порталов открытых данных, созданный в Open Knowledge Foundation. CKAN позиционируется как система управления данными (Data Management System) и поддерживается компанией Datatopian, созданной командной Open Knowledge Foundation. На базе CKAN работают порталы открытых данных США (data.gov), Австралии (data.gov.au) и ещё десятков стран.

Лучшие практики использования: CKAN применяется практически во всех продвинутых порталах открытых данных по всему миру, и активно расширяется пользователями под особенности и специфику публикуемых данных. При этом код CKAN довольно сложен, его развертывание требует серьёзных технических навыков, а эксплуатация портала требует больших ресурсов, чем в случае альтернативных продуктов. Поэтому CKAN рекомендуется к использованию тем, кто понимает, как и зачем создается портал/ каталог открытых данных.

DKAN

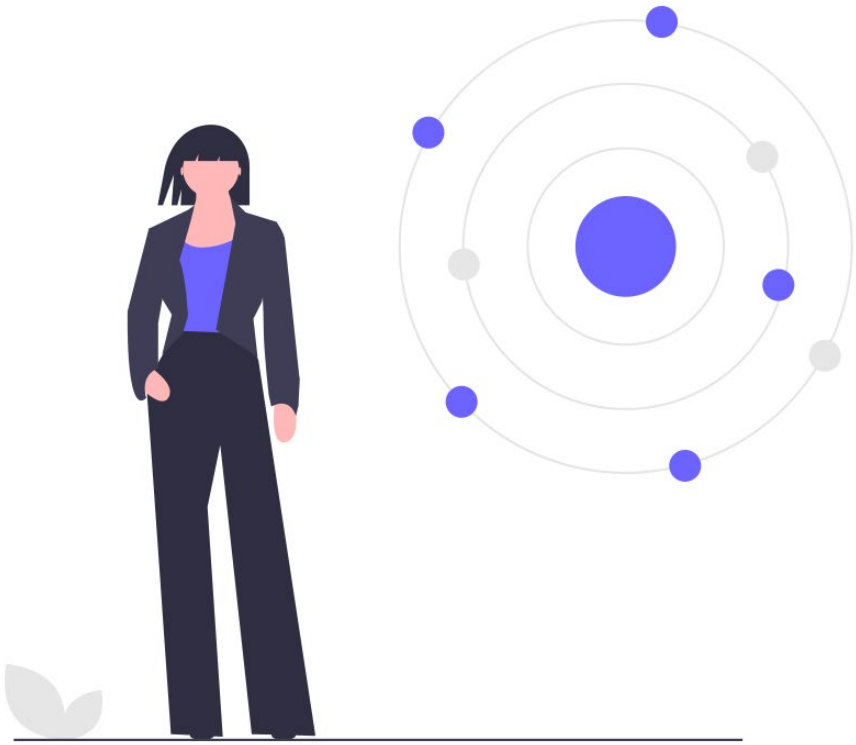
Назначение инструмента	Каталоги и порталы данных
Уровень сложности	Продвинутый
Доступен исходный код	Да

Тип продукта	Веб-приложение
Ссылка на сайт	getdkan.org
Ссылка на открытый код	github.com/getdkan/dkan
Лицензия на код	GNU Affero General Public
Язык разработки	PHP
Поддерживаемые типы данных	Любые типы данных

Каталог открытых данных на базе продукта Drupal 8 создан как более простая альтернатива CKAN. Продукт поддерживается НКО CivicAction. На базе DKAN работает портал открытых данных Российской Федерации https://data.gov.ru

Лучшие практики использования: DKAN гораздо проще других каталогов открытых данных, поэтому его наиболее эффективно применяют те, кто ищет вариант быстрого создания портала открытых данных, без сложных функций и возможностей, присущих, например, CKAN. Также DKAN является более простым в технической поддержке из-за того, что разработчиков на языках программирования PHP/Drupal значительно больше, чем других разработчиков.

Сбор и извлечение данных



Обработка данных включает разнообразное число операций по преобразованию данных из одного формата в другой, приведению их к пригодности для последующего анализа, визуализации или передачи на хранение. Обработка данных включает также процессы очистки

данных и их стандартизации, необходимые для подготовки данных к публикации.

Для кого это важно: Владельцы данных, пользователи данных

Сложность использования: Продвинутая

Рекомендации: У каждого из инструментов сбора и извлечения данных свои функции, часто уникальные и не воспроизводимые другими способами. Необходимо использовать те инструменты, которые необходимы для решений вашей конкретной задачи.

Tabula

Назначение инструмента	Сбор и извлечение данных, Обработка данных
Уровень сложности	Сложный
Доступен исходный код	Да
Тип продукта	Веб приложение
Ссылка на сайт	github.com/tabulapdf/tabula
Ссылка на открытый код	github.com/tabulapdf/tabula
Лицензия на код	MIT License
Язык разработки	Python
Поддерживаемые типы данных	PDF, CSV

Tabula — это программная библиотека и утилита командной строки (CLI) для извлечения таблиц из файлов PDF. Tabula используется для автоматизации потоковой обработки PDF документов и является наиболее популярным продуктом для выполнения этой функции.

Лучшие практики использования: Tabula является почти безальтернативным из бесплатных и открытых инструментов для обработки таблиц из PDF документов. Альтернативны ему только коммерческие инструменты, часто показывающие лучшее качество распознавания, но требующие существенных расходов. Поэтому в работе с табличными PDF документами выбор Tabula является наиболее очевидным.

PhantomJS

Назначение инструмента	Сбор и извлечение данных, Обработка данных
Уровень сложности	Сложный
Доступен исходный код	Да
Тип продукта	Командная строка
Ссылка на сайт	phantomjs.org
Ссылка на открытый код	github.com/ariya/phantomjs

Лицензия на код	BSD-3-Clause License
Язык разработки	C
Поддерживаемые типы данных	HTML

PhantomJS — это программная библиотека и инструмент командной строки для имитации работы браузера. PhantomJS используется для задач скрейпинга, извлечения данных из веб-страниц.

Лучшие практики использования: PhantomJS используется в задачах, связанных со сбором данных с веб страниц. За счёт имитации поведения пользователя этот инструмент позволяет обманывать системы защиты от сбора данных и выгружать данные с сайтов в максимально короткие сроки.

Metawarc

Назначение инструмента	Сбор и извлечение данных, Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Командная строка
Ссылка на сайт	github.com/datacoon/metawarc
Ссылка на открытый код	github.com/datacoon/metawarc
Лицензия на код	MIT License
Язык разработки	Python
Поддерживаемые типы данных	WARC

Metawarc — это утилита командной строки, применимая для извлечения метаданных из файлов веб-архивов, WARC файлов. Инструмент используется для сбора данных в таких проектах, как международный веб-архив (archive.org) или Национальный цифровой архив России (ruarxive.org).

Лучшие практики использования: Metawarc позволяет в короткие сроки извлекать метаданные из WARC файлов, и альтернативой ему является только использование инструментов и программных библиотек языков программирования для работы с WARC файлами.

Очистка и контроль качества данных

Обработка данных включает также процессы очистки данных и их стандартизации, необходимые для подготовки данных к публикации. Очистка и контроль качества — обязательные операции после получения данных, в процессе их обработки, при планировании их сбора и во многих других случаях.

Для кого это важно: Владельцы данных, пользователи данных

Сложность использования: Продвинутая

Рекомендации: Data Cleaner — это базовый выбор для всех, кто хочет обойтись без программирования. В остальных случаях предпочтительнее использовать такие инструменты, как great expectations — инструмент контроля качества данных по мере их поступления.

Data Cleaner

DataCleaner

Назначение инструмента	Очистка и контроль качества данных
Уровень сложности	Простой
Доступен исходный код	Да

Тип продукта	Командная строка
Ссылка на сайт	datacleaner.org
Ссылка на открытый код	github.com/datacleaner/DataCleaner
Лицензия на код	LGPL-3.0 License
Язык разработки	Java
Поддерживаемые типы данных	SQL

Data Cleaner — это настольное приложение с открытым кодом для задач очистки данных. Приложение подключается напрямую к базе данных, поэтому все изменения/исправления также сохраняются в базе данных.

Лучшие практики использования: Data Cleaner является лучшим из бесплатных инструментов по очистке данных непосредственно в месте их хранения, в СУБД. Альтернативы ему — это только программирование, использование инструментов построения запросов (query builders) или коммерческих инструментов очистки данных от ведущих производителей.

Курсы и лекции

Наименование	Ссылка	Уровень сложности	Автор
Семинар «Анализ открытых данных для НКО. Урок 2: чистим и структурируем данные»	youtu.be/lE26paEZXLc	Простой	Иван Бегтин

Great Expectations

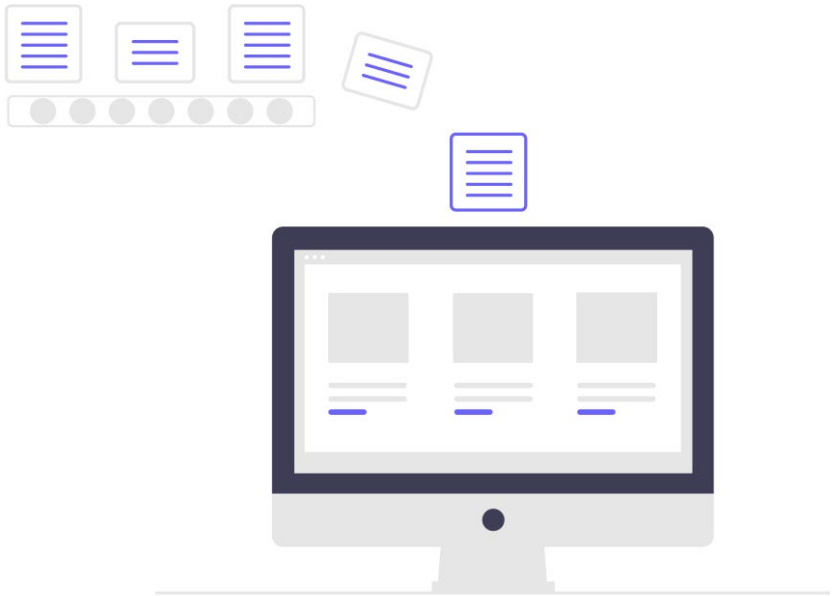
Назначение инструмента	Очистка и контроль качества данных
Уровень сложности	Сложный
Доступен исходный код	Да
Тип продукта	Командная строка
Ссылка на сайт	greatexpectations.io
Ссылка на открытый код	github.com/great-expectations/great_expectations
Лицензия на код	Apache-2.0 License
Язык разработки	Python

Поддерживаемые типы данных	CSV

Great Expectations — это специализированная библиотека для языка Python, используемая для контроля качества данных. Она работает с табличными данными и позволяет оценить качество поступающих табличных данных (обычно в CSV формате) на предмет наличия аномалий и иных расхождений с предварительно подготовленными правилами. Используется такими компаниями как Video, Heineken и многими другими.

Лучшие практики использования: Great Expectations является инструментом обеспечения и контроля качества данных в процессе их обработки. Ключевой способ его использования — включение данных в трубы данных (data pipelines), используемые на этапах передачи их из источника данных к итоговому результату.

Обработка данных



Обработка данных включает разнообразное число операций по преобразованию данных из одного формата в другой, приведению их к пригодности для последующего анализа, визуализации или передачи на хранение. Обработка данных включает также процессы очистки данных и их стандартизации, необходимые для подготовки данных к публикации.

Для кого это важно: Владельцы данных, пользователи данных

Сложность использования: Продвинутая

Рекомендации: Для работы без программирования наиболее оптимален OpenRefine, в задачах, требующих интенсивной программной обработки, используют такие языки разработки как Python, в промежуточных задачах и под определенные форматы файлов, другие инструменты.

OpenRefine

Назначение инструмента	Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	openrefine.org
Ссылка на открытый код	github.com/OpenRefine/OpenRefine
Лицензия на код	BSD-3-Clause License
Язык разработки	Python
Поддерживаемые типы данных	CSV, XML, JSON, JSON-LD, N3, N-Triples, Turtle, RDF/XML, XLSX, XLS, ODS, text, MARC, PC-Axis(PX)

OpenRefine — это специальное приложение (работает как локально, так и на собственном сервере) для очистки и обработки данных. Изначально OpenRefine создали в компании Google под названием Google Refine, затем передали его в открытый доступ в виде открытого кода и свободного проекта, как OpenRefine. Поддерживает большое число форматов файлов на импорт и на экспорт, обработку данных по колонкам, использование скриптов Python и GREL для потоковой построчной обработки и возможность ручной очистки данных.

Лучшие практики использования: OpenRefine наиболее эффективно применяется для задач ручной и полуручной проверки и очистки данных применительно к каждой колонке табличных данных. Для работы OpenRefine не требуется знаний программирования, но при их наличии можно значительно автоматизировать свою работу.

Курсы и лекции

Наименование	Ссылка	Уровень сложности	Автор
OpenRefine для работы с финансовыми данными	youtu.be/Z2Qlc78tEG8	Простой	Ольга Пархимович
Семинар «Анализ открытых данных для НКО. Урок 2: чистим и структурируем данные»	youtu.be/lE26paEZXLc	Простой	Иван Бегтин

CSVKit

Назначение инструмента	Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Командная строка
Ссылка на сайт	csvkit.rtfd.org
Ссылка на открытый код	github.com/wireservice/csvkit
Лицензия на код	MIT License
Язык разработки	Python
Поддерживаемые типы данных	CSV

CSVKit — это набор утилит командной строки, позволяющих производить и простые, и сложные манипуляции над табличными данными в формате CSV.

Лучшие практики использования: CSVKit — это наиболее очевидный набор инструментов для тех, кто постоянно работает с командной строкой и вынужден работать с CSV файлами большого объёма. В этом случае CSVkit позволяет быстро разделять файлы, объединять их, фильтровать и преобразовывать. Многие из этих операций можно выполнять и другими инструментами, такими как grep, sed и т.д., но это потребует больше времени и усложнит команды.

Undatum

Назначение инструмента	Обработка данных
Уровень сложности	Сложный
Доступен исходный код	Да
Тип продукта	Командная строка
Ссылка на сайт	github.com/datacoon/undatum
Ссылка на открытый код	github.com/datacoon/undatum
Лицензия на код	MIT License
Язык разработки	Python
Поддерживаемые типы данных	CSV, JSON, BSON, JSON lines

Undatum — это утилита командной строки для преобразования данных большого объёма в форматах CSV, JSON и BSON. Утилита создавалась для обработки данных проекта «Госрасходы» (spending.gov.ru) и обработки дампов данных, опубликованных в проекте.

Лучшие практики использования: Undatum подходит для работы с BSON файлами и файлами JSON lines. Этот инструмент является переносом возможностей таких инструментов, как Csvkit, к файлам более сложной структуры и большего объёма.

Python

Назначение инструмента	Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Командная строка
Ссылка на сайт	python.org
Ссылка на открытый код	github.com/python/cpython
Лицензия на код	Python Software Foundation License 2
Язык разработки	C
Поддерживаемые типы данных	Любые типы и форматы данных

Python — это универсальный язык программирования, получивший широкое распространение и применение в работе с данными. Сочетая гибкие возможности языка и большое число инструментов, созданных с его помощью, его часто используют для задач сбора, обработки, анализа и визуализации данных.

Лучшие практики использования: Python применим во всех задачах, в которых не справляются типовые/стандартные инструменты обработки данных. С помощью скриптов на Python легко написать код по преобразованию любых данных.

Курсы и лекции

Наименование	Ссылка	Уровень сложности	Автор
COVID-19: Как смоделировать распространение коронавируса? Воркшоп по анализу данных	youtu.be/-u3nPFdw2UQ	Продвинутый	Дмитрий Сергеев
Мастер-класс «О чем говорят депутаты Госдумы? Анализ текстовых данных на Python»	youtu.be/8IaE-e8kjd8	Продвинутый	Дмитрий Сергеев

Apache Hadoop

Назначение инструмента	Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	CLI
Ссылка на сайт	hadoop.apache.org
Ссылка на открытый код	github.com/apache/hadoop hadoop.apache.org/docs/r3.2.2/
Лицензия на код	MIT License
Язык разработки	Java, C++
Поддерживаемые типы данных	Любые типы и форматы данных

Программная библиотека Apache Hadoop представляет собой фреймворк, позволяющий распределённо обрабатывать большие массивы данных по кластерам компьютеров с использованием простых моделей программирования. Она предназначена для масштабирования с отдельных серверов до тысяч машин, каждый из которых предлагает локальные вычисления и хранение.

Лучшие практики использования: Библиотека предназначена для обнаружения и обработки отказов на прикладном уровне, поэтому она предоставляет высокодоступный сервис поверх кластера компьютеров, каждый из которых может быть подвержен сбоям.

Natasha

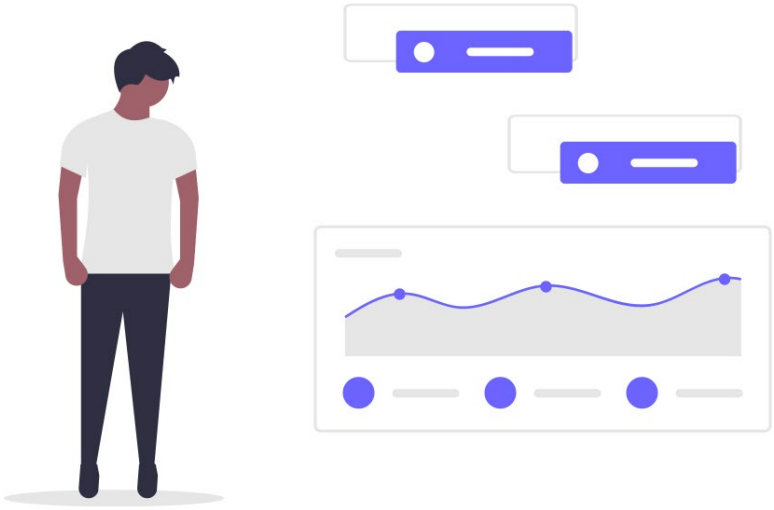
Назначение инструмента	Обработка текстовых данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Библиотеки для Python
Ссылка на сайт	natasha.github.io
Ссылка на открытый код	github.com/natasha/natasha
Лицензия на код	MIT License
Язык разработки	Python

Поддерживаемые типы данных	

Проект Natasha — набор Python-библиотек для обработки текстов на естественном русском языке. Natasha решает основные задачи NLP для русского языка: токенирование, сегментация предложения, вставка слова, теги морфологии, лемматизация, нормализация фразы, синтаксический разбор, NER-тегирование, извлечение фактов.

Лучшие практики использования: Для новостных статей качество исполнения на всех задачах сравнимо или превосходит существующие решения. Качество по каждой задаче аналогично или лучше, чем по текущим SOTA для русского языка в новостях, см. раздел «Оценка». Библиотека поддерживает Python 3.5+ и PyPy3, не требует GPU, зависит только от NumPy.

Аналитика и Business Intelligence



Аналитическая работа — это одна из наиболее очевидных областей применения данных и их наглядного представления для последующего принятия решений. Аналитические инструменты включают как относительно простые инструменты, так и продвинутые научные записные книжки.

Для кого это важно: Пользователи данных

Сложность использования: Продвинутая

Рекомендации: Metabase — идеальный инструмент в случае, если у результатов анализа данных должны быть внешние пользователи. Jupyter Notebook и Orange — наиболее эффективны для внутренней работы команд дата-аналитиков и обучения студентов.

Metabase

Назначение инструмента	Аналитика, Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб приложение
Ссылка на сайт	metabase.com
Ссылка на открытый код	github.com/metabase/metabase
Лицензия на код	GNU Affero General Public License
Язык разработки	Java
Поддерживаемые типы данных	CSV, SQL

Metabase — аналитический продукт для построения панелей управления с интеграцией более чем с 20 базами данных (Postgres, Oracle, MS SQL и другие). Опубликован с открытым кодом и пригоден для быстрого развертывания и начала работы.

Лучшие практики использования: Metabase можно назвать самым очевидным решением для задач, связанных с быстрым построением дашбордов из предварительно подготовленных баз данных. В нем нет функций очистки и обработки данных, так как этот инструмент предполагает, что все задачи подготовки данных уже решены. Metabase даёт возможность визуального представления данных в различных формах.

Jupyter Notebook

Назначение инструмента	Аналитика, Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Командная строка
Ссылка на сайт	jupyter.org
Ссылка на открытый код	github.com/jupyter/notebook
Лицензия на код	Modified BSD License
Язык разработки	Python

Поддерживаемые типы данных	CSV, JSON, XML

Jupyter Notebook — это инструмент работы с цифровыми записными книжками, используемыми исследователями в области работы с данными. Jupyter Notebook используется тысячами команд разработчиков и аналитиков по всему миру, в большинстве крупнейших компаний, работающих с данными.

Лучшие практики использования: это стандарт де-факто для всех исследователей, работающих с данными. В задачах обмена практиками, работы с данными и обучения студентов используется чаще всего именно Jupyter Notebook.

Orange

Назначение инструмента	Аналитика, Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб приложение
Ссылка на сайт	orangedatamining.com
Ссылка на открытый код	github.com/biolab/orange3
Лицензия на код	GNU General Public License
Язык разработки	Python
Поддерживаемые типы данных	CSV, SQL

Orange — это инструмент для визуализации и добычи данных, пригодный для специалистов любой квалификации. Orange не требует программирования или глубоких математических знаний и предоставляет интерфейс для удобной обработки данных и визуализации.

Лучшие практики использования: исторически Orange появился раньше других аналитических инструментов по работе с данными, и во многих ВУЗах накоплен большой опыт его изучения и использования. Он наиболее применим там, где есть причины и основания использования именно настольных, а не веб-приложений. Также Orange подходит для обучения работе с данными тех, кто не имеет опыта программирования.

Курсы и лекции

Наименование	Ссылка	Уровень сложности	Автор
Дата-среда: «Интерактивный data mining»	youtu.be/i1PkfRbGx6s	Простой	Дмитрий Стефановский

Querybook

Назначение инструмента	Аналитика, Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	www.querybook.org
Ссылка на открытый код	github.com/pinterest/querybook
Лицензия на код	Apache-2.0 License
Язык разработки	Python
Поддерживаемые типы данных	SQL

Querybook — система построения, сохранения и обмена запросами к базам данных, позволяющая быстро формировать аналитические выборки с помощью SQL. Создана и передана в открытый доступ с открытым кодом компанией Pinterest в 2021 году.

Лучшие практики использования: QueryBook — это новый инструмент, похожий на бесплатную и открытую реализацию других построителей запросов таких, как trevor.io и его аналогов, позволяющих быстро визуализировать запросы к базам данных и подключающиеся к почти любым СУБД.

Knime Analytics Platform

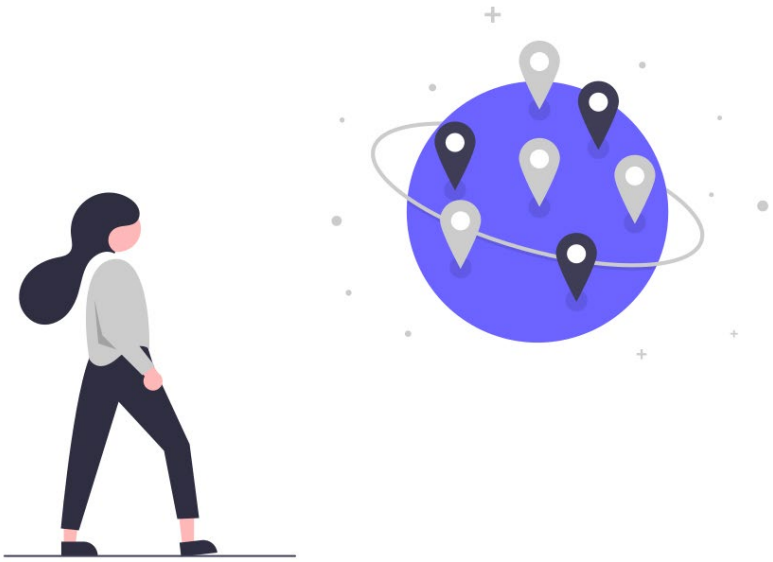
Назначение инструмента	Анализ данных (data science)
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	knime.com
Ссылка на открытый код	docs.knime.com
Лицензия на код	Open Source GPLv3
Язык разработки	Java

Поддерживаемые типы данных	PDF и структурированные типы данных (CSV, XLS, JSON, XML и др.), неструктурированные типы данных (изображения, документы, сети, молекулы и др.) или данные временных рядов.

Аналитическая платформа KNIME Analytics — это программное обеспечение с открытым исходным кодом для анализа данных (data science).

Лучшие практики использования: В отличие от других продуктов с открытым исходным кодом, KNIME Analytics Platform не имеет ограничений на среду исполнения или размер данных. При наличии необходимого локального или облачного пространства и вычислительной мощности, вы можете запускать проекты с действительно большими объемами данных (более миллиарда строк).

Работа с геоданными



Работа с геоданными имеет довольно много особенностей, связанных со спецификой геокодирования, потребности в специальных инструментах и практиках, сформировавшихся у сообществ геоинформатики.

Для кого это важно: Владельцы данных, пользователи данных

Сложность использования: Продвинутая

Рекомендации: В случае геоданных выбор инструмента с открытым исходными кодом очевиден — это QGIS, который представляет стандарт работы с геоданными с большим функционалом.

QGIS

Назначение инструмента	Работа с геоданными
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	qgis.org
Ссылка на открытый код	github.com/qgis/QGIS
Лицензия на код	GNU General Public License
Язык разработки	Python
Поддерживаемые типы данных	CSV, JSON, GeoJSON

QGIS — это продвинутый инструмент подготовки, обработки, визуализации и анализа данных в многочисленных форматах данных, существующих в области геопространственного анализа.

Лучшие практики использования: у QGIS практически нет бесплатных альтернатив, а сам он является альтернативой для платного ArcGIS — инструмента для работы с геоданными от одноимённой компании.

Курсы и лекции

Наименование	Ссылка	Уровень сложности	Автор
Дата-среда: город и пространственные данные	youtu.be/Vu6t04-yZSY	Продвинутый	Егор Котов

Unfolded.ai

Назначение инструмента	Работа с геоданными, визуализация
Уровень сложности	Простой
Доступен исходный код	Да

Тип продукта	Веб-приложение
Ссылка на сайт	unfolded.ai
Ссылка на открытый код	docs.unfolded.ai/development
Лицензия на код	github.com/UnfoldedInc/unfolded.gl#unfoldedgl
Язык разработки	C++, Python
Поддерживаемые типы данных	JSON, HTML

Unfolded — это платформа для геопространственного анализа. Она включает набор инструментов для работы с большими геопространственными данными. Решает задачи визуализации больших данных о местоположении, геопространственного анализа и др.

Лучшие практики использования: Открытый программный фреймворк, используемый для создания Unfolded Studio, предлагает возможности геопространственной визуализации и анализа на GPU, обеспечивая графику и производительность. В отличие от других открытых геопространственных проектов, эти фреймворки спроектированы с нуля и не являются монолитными, предлагая композитные и многократно используемые компоненты, которые можно выбирать в соответствии со всеми типами разработки геопространственных приложений.

Курсы и лекции

Наименование	Ссылка	Уровень сложности	Автор
Мастер-класс. «Создание карт без специального программного обеспечения»	youtu.be/iwXWGcNwRbU	Простой	Татьяна Балтыжакова

GraphHopper's

Назначение инструмента	Инструмент для решения проблем с маршрутизацией транспортных средств
Уровень сложности	Простой
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	graphhopper.com
Ссылка на открытый код	github.com/graphhopper

Лицензия на код	Apache License 2.0
Язык разработки	Java, JavaScript
Поддерживаемые типы данных	XML, OSM, pbf

GraphHopper — это быстрый и экономичный по памяти движок маршрутизации Java, выпущенный под лицензией Apache 2.0. По умолчанию он использует данные OpenStreetMap и GTFS, но может импортировать и другие источники данных. Предоставляет простой веб-интерфейс API, включая JavaScript и Java-клиенты.

Лучшие практики использования: Инструмент основан на Java. Работает с OpenStreetMap (osm/xml и pbf) и может быть адаптирован под пользовательские данные. Интеграция OpenStreetMap: хранит и учитывает тип дороги, ограничение скорости, покрытие, барьеры, ограничения доступа, паромы, ограничения условного доступа и др.

Overpass Turbo

Назначение инструмента	Инструмент интеллектуального анализа данных для OpenStreetMap
Уровень сложности	Простой
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	overpass-turbo.eu
Ссылка на открытый код	github.com/tyrasd/overpass-turbo
Лицензия на код	MIT License
Язык разработки	JavaScript
Поддерживаемые типы данных	JSON

Overpass Turbo — это веб-утилита для поиска данных OpenStreetMap.

Лучшие практики использования: С Overpass Turbo можно выполнять Overpass API запросы и анализировать результаты из OSM на интерактивной карте.

Openroute Service

openroute service

Назначение инструмента	Инструмент по работе с геоданными OSM
Уровень сложности	Простой
Доступен исходный код	Да

Тип продукта	Веб-приложение
Ссылка на сайт	hopenrouteservice.org
Ссылка на открытый код	github.com/GIScience/openrouteservice
Лицензия на код	GNU General Public License 3.0
Язык разработки	Java
Поддерживаемые типы данных	XML, CSV

API Open Route Service предоставляет глобальные пространственные сервисы, потребляя бесплатные географические данные, генерируемые пользователями и собираемые ими совместно, непосредственно из OpenStreetMap.

Лучшие практики использования: Поддерживается Гейдельбергским университетом, что дает команде сервиса преимущество для разработки собственных алгоритмов и использования передовых технологий с открытым исходным кодом в области геопространственных данных.

deck.gl

Назначение инструмента	Визуальный анализ больших объемов геоданных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	deck.gl
Ссылка на открытый код	github.com/visgl/deck.gl
Лицензия на код	MIT License
Язык разработки	JavaScript, Python
Поддерживаемые типы данных	JSON

Deck.gl — это фреймворк на базе WebGL-визуалиации больших массивов данных. Deck.gl отображает данные (обычно массив JSON-объектов) в стек визуальных слоев: например, иконки, полигоны, тексты; и смотрит на них с помощью представлений: например, в виде карты. Deck.gl спроектирован так, чтобы быть легко настраиваемым. Все слои поставляются с гибкими API, позволяющими программно управлять каждым аспектом рендеринга. Все основные классы легко расширяются пользователями для решения пользовательских задач.

Лучшие практики использования: Пользователи могут быстро получить визуальные результаты с минимальными усилиями, создавая существующие слои или используя расширяемую архитектуру deck.gl для удовлетворения индивидуальных потребностей.

kepler.gl

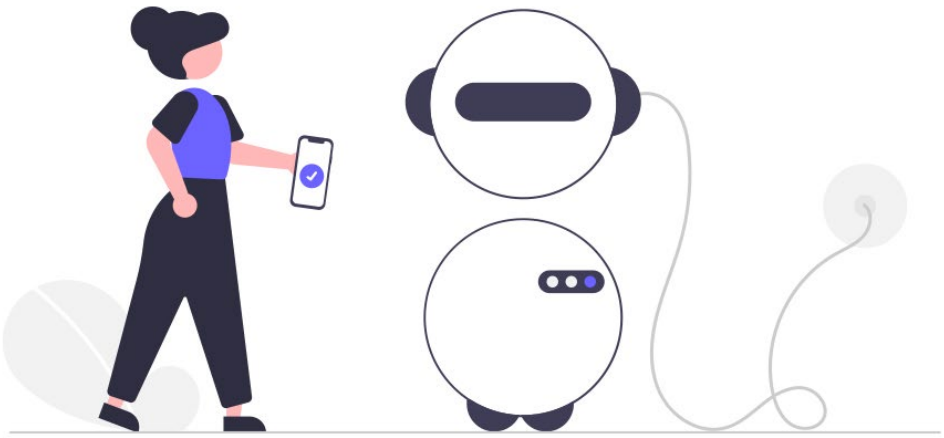
kepler.gl

Назначение инструмента	Инструмент геопространственного анализа с открытым исходным кодом для крупномасштабных массивов данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	kepler.gl
Ссылка на открытый код	github.com/keplergl/kepler.gl
Лицензия на код	MIT License
Язык разработки	JavaScript
Поддерживаемые типы данных	JSON

Kepler.gl — это диагностическое, высокопроизводительное веб-приложение для визуальной разведки крупномасштабных геолокационных массивов данных. Построенный на основе Mapbox GL и deck.gl, kepler.gl может визуализировать миллионы точек, представляющих тысячи перемещений, и быстро выполнять пространственные агрегации.

Лучшие практики использования: Kepler.gl также является компонентом React, который использует Redux для управления своим состоянием и потоком данных. Он может быть встроен в другие приложения React-Redux и обладает высокой настраиваемостью.

Машинное обучение и искусственный интеллект



Технологии машинного обучения и искусственного интеллекта активно используются для

создания автоматизированных алгоритмов и систем поддержки принятия решений. Среди этих технологий немало тех, кто публикует открытый код, что позволяет разрабатывать и раскрывать собственные решения.

Для кого это важно: Пользователи данных

Сложность использования: Сложная

Рекомендации: Fairlearn — один из немногих инструментов с открытым кодом, позволяющий повысить честность алгоритмов ИИ.

Fairlearn

Назначение инструмента	Аналитика, Обработка данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб приложение
Ссылка на сайт	fairlearn.org
Ссылка на открытый код	github.com/fairlearn/fairlearn
Лицензия на код	MIT License
Язык разработки	Python
Поддерживаемые типы данных	CSV, SQL, Dataframe

FairLearn — это программная библиотека и специальное расширение для Jupyter Notebook, позволяющее исследователям, работающим с данными, повышать справедливость и честно алгоритмов ИИ.

Лучшие практики использования: FairLearn является очевидным выбором для тестирования алгоритмов исследователями, работающими в области ИИ. Его использование позволит проверить честность алгоритмов и использовать лучшие практики в этой области.

BigARTM

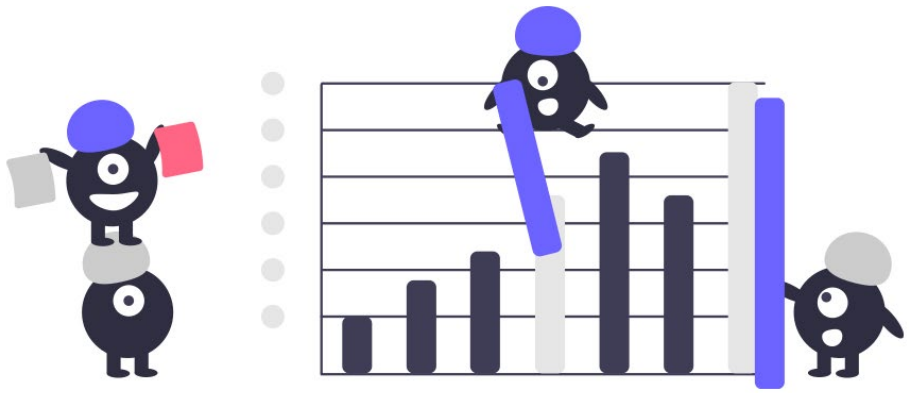
Назначение инструмента	Аналитика, Обработка данных
Уровень сложности	Продвинутый

Доступен исходный код	Да
Тип продукта	Библиотека
Ссылка на сайт	github.com/bigartm/bigartm
Ссылка на открытый код	github.com/bigartm/bigartm
Лицензия на код	New BSD License
Язык разработки	Python, C++
Поддерживаемые типы данных	TXT

BigARTM — это инструмент тематического моделирования, основанный на новой методике под названием Additive Regularization of Topic Models (аддитивная регуляризация тематических моделей). Эта техника эффективно строит мульти-объективные модели, добавляя к критерию оптимизации взвешенные суммы регуляризаторов.

Лучшие практики использования: BigARTM сочетает в себе очень разные задачи, в том числе спарринг, сглаживание, декорирование тем и многие другие. Такое сочетание регуляризаторов значительно улучшает качественные показатели семантического анализа.

Визуализация данных



Визуализация данных позволяет максимально простым и наглядным образом представить данные или результаты анализа данных для широкой аудитории. Задача визуализации данных заключается не только в том, чтобы представить данные визуально, а в том, чтобы передать основную идею, передать сообщение, рассказать аудитории историю, скрытую в данных.

Для кого это важно: дизайнерам, аналитикам, руководителям, студентам и исследователям.

Сложность использования: разные уровни сложности, от простого к сложному.

Рекомендации: Визуализация данных процесс очень индивидуальный, выбрать какой-то

конкретный инструмент можно исходя из задач. Для большей части типовых задач по визуализации подойдет сервис Datawrapper, который не требует специальных навыков программирования и знаний основ дизайна. Как универсальный инструмент, можно порекомендовать D3J — достаточно гибкий, но и не самый простой в использовании.

RAWGraphs

RAWGraphs

Назначение инструмента	Инструмент для визуализации данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	rawgraphs.io
Ссылка на открытый код	github.com/rawgraphs/rawgraphs-app
Лицензия на код	Apache License 2.0
Язык разработки	JavaScript (NodeJS)
Поддерживаемые типы данных	CSV, TSV, SVG, copied-and-pasted texts from other applications

RAWGraphs — веб-приложение с открытым исходным кодом для создания статических визуализаций данных, которые предназначены для дальнейшей модификации.

Лучшие практики использования: RAWGraphs позволяет решать задачи, недоступные в решениях других инструментов визуализации. Приложение подходит для создания сложных визуализаций.

Datawrapper

Datawrapper

Назначение инструмента	Инструмент для визуализации данных
Уровень сложности	Простой
Доступен исходный код	Да
Тип продукта	Веб-приложение
Ссылка на сайт	datawrapper.de
Ссылка на открытый код	github.com/datawrapper

Лицензия на код	MIT
Язык разработки	PHP, JavaScript
Поддерживаемые типы данных	CSV, XLS(X), copied-and-pasted texts from other applications

Datawrapper — платформа визуализации данных с открытым исходным кодом, помогающая каждому создавать простые, корректные и встраиваемые графики за считанные минуты. Большая часть функциональности в Datawrapper обеспечивается плагинами, некоторые из них имеют открытый исходный код.

Лучшие практики использования: Инструмент имеет все базовые функции, необходимые для создания визуализаций для статей, отчетов или публикаций. Чтобы пользоваться Datawrapper, не нужно иметь навыков программирования и дизайна.

Курсы и лекции

Наименование	Ссылка	Уровень сложности	Автор
Семинар «От данных к истории: подготовка исследований к публикации в медиа»	youtu.be/OkM_x8Rqerc	Простой	Алексей Смагин

Gephi

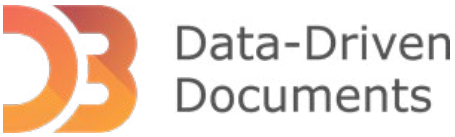
Назначение инструмента	Инструмент для визуализации связей в данных (графов)
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Настольное приложение
Ссылка на сайт	gephi.org
Ссылка на открытый код	github.com/gephi/gephi
Лицензия на код	CDDL 1.0 and GNU General Public License v3 gephi.org/developers/license/
Язык разработки	Java
Поддерживаемые типы данных	GML file, GEFX file и другие

Gephi — это интерактивная платформа для визуализации и исследования всех видов сетей и сложных систем, динамических и иерархических графов. Инструмент имеет открытый исходный код. Благодаря встроенному движку OpenGL, Gephi может работать с очень большими сетями. Может визуализировать сети до миллиона элементов.

Лучшие практики использования: Цель Gephi — помочь аналитикам данных строить гипо-

тезы, интуитивно обнаруживать закономерности, изолировать особенности структуры или сбои во время поиска данных. Это инструмент, дополняющий традиционную статистику, так как визуальное мышление с интерактивными интерфейсами признано средством, упрощающим анализ данных.

D3JS



Назначение инструмента	Инструмент для визуализации данных
Уровень сложности	Продвинутый
Доступен исходный код	Да
Тип продукта	Настольное приложение
Ссылка на сайт	d3js.org
Ссылка на открытый код	github.com/d3
Лицензия на код	BSD 3-Clause «New» or «Revised» License github.com/d3/d3/blob/master/LICENSE
Язык разработки	JavaScript
Поддерживаемые типы данных	JSON, SVG, CSV, TSV, XML, HTML

D3.js — это библиотека JavaScript для работы с данными. D3 помогает визуализировать данные с помощью HTML, SVG и CSS. Акцент D3 на веб-стандартах дает вам все возможности современных браузеров, не привязывая себя к проприетарному фреймворку, сочетая в себе мощные компоненты визуализации и основанный на данных подход к манипуляциям с DOM.

Лучшие практики использования: С помощью библиотеки D3JS можно визуализировать данные в разных типах графиков и диаграмм, представлять данные в анимированном формате, создавать интерактивные визуализации, а также D3 включает в себя инструменты для количественного анализа данных.

Итоговые рекомендации

Подготовка и упаковка данных

DataCurator — для работы при недостатке технических навыков, и Frictionless Data — для задач, требующих автоматизации подготовки данных.

Создание порталов и каталоги открытых данных

CKAN обладает наиболее продвинутыми возможностями по разработке порталов открытых данных. Это лучший выбор, если есть техническая возможность его развернуть. DKAN выступает в роли более простой альтернативы, удобной в условиях дефицита технических ресурсов.

Сбор и извлечение данных

У каждого из инструментов сбора и извлечения данных свои функции, часто уникальные и не воспроизводимые другими способами. Необходимо использовать те инструменты, которые необходимы для решений вашей конкретной задачи.

Очистка и контроль качества данных

Data Cleaner — это базовый выбор для всех, кто хочет обойтись без программирования, в остальных случаях — предпочтительнее использовать такие инструменты, как great expectations — приложение контроля качества данных по мере их поступления.

Обработка данных

Для работы без программирования наиболее оптимален — OpenRefine. В задачах, требующих интенсивной программной обработки, используют такие языки разработки как Python. В промежуточных задачах и под определенные форматы файлов — другие инструменты.

Аналитика и Business Intelligence

Metabase — идеальный инструмент в случае, если у результатов аналитики должны быть внешние пользователи. Jupyter Notebook и Orange — наиболее эффективны для внутренней работы команд дата-аналитиков и обучения студентов.

Машинное обучение и искусственный интеллект

Fairlearn — инструмент, позволяющий повысить честность алгоритмов искусственного интеллекта. Одно из немногих решений с открытым кодом в этой области.

Визуализация данных

Визуализация данных процесс очень индивидуальный, выбрать какой-то конкретный инструмент можно исходя из задач. Для большей части типовых задач по визуализации подойдет сервис Datawrapper, который не требует специальных навыков программирования и знаний основ дизайна. Как универсальный инструмент, можно порекомендовать D3J — достаточно гибкий, но и не самый простой в использовании.