УДК 004.622

Быков Борис Евгеньевич, студент, кафедра информационных систем и технологий, Санкт-Петербургский государственный университет аэрокосмического приборостроения, г. Санкт-Петербург

ОБЗОР ОТКРЫТЫХ ИСТОЧНИКОВ ДАТАСЕТОВ ДЛЯ НАУЧНЫХ ИССЛЕДОВАНИЙ

Аннотация. В статье рассматриваются различные открытые источники готовых датасетов. В условиях развития цифровых технологий, данные стали важным ресурсом для научных исследований. Сбор и подготовка датасетов для последующего анализа - долгий и кропотливый труд. Открытые источники позволяют исследователям сосредоточиться на решении прикладных задач без необходимости первичной обработки информации. Использование подобных ресурсов способствует развитию междисциплинарных проектов и ускоряет процесс получения результатов. В статье рассмотрены как специализированных платформ для хранения и обмена подготовленными данными, так и научные репозитории.

Annotation. The article examines various open sources of ready-to-use datasets. In the context of digital technology development, data has become an important resource for scientific research. Collecting and preparing datasets for subsequent analysis is a time-consuming and labor-intensive process. Open sources allow researchers to focus on solving applied problems without the need for primary data processing. The use of such resources facilitates the development of interdisciplinary projects and accelerates the results achievement process. The article overviews both specialized platforms for storing and sharing prepared data and scientific repositories.

Ключевые слова: наука о данных, датасеты, открытые данные, источники данных, обзор

Keywords: data science, datasets, open data, data sources, overview

В цифровую эпоху сфера больших данных [1, 2] вызывает стремительно растущий интерес. Данные стали ключевым ресурсом для исследований, принятия решений и разработки алгоритмов машинного обучения. Однако их сбор и подготовка [3] – трудоемкие процессы, которые требуют значительных временных и технических затрат. По этой причине открытые датасеты приобретают особую ценность и существует целый ряд их источников, позволяя исследователям, аналитикам и разработчикам сосредоточиться на решении прикладных задач, минуя процесс первичной обработки данных.

Первое, о чем стоит упомянуть, – это **специализированные платформы**, которые предназначены для поиска необходимых наборов данных. Один из наиболее известных и крупных из них – платформы **Kaggle**. На ней содержится огромная коллекция самых разнообразных датасетов, которая продолжает расти за счет сообщества – любой желающий в праве поделиться своими данным с другими людьми. Также платформа регулярно проводит различного рода соревнования, подогревающие интерес к теме больших данных. Многие наборы данных на Kaggle сопровождаются примерами кода и визуализацией. Кроме того, всегда можно прочитать комментарии к тому или иному датасету, чтобы лучше понять подходит ли он под конкретную задачу.

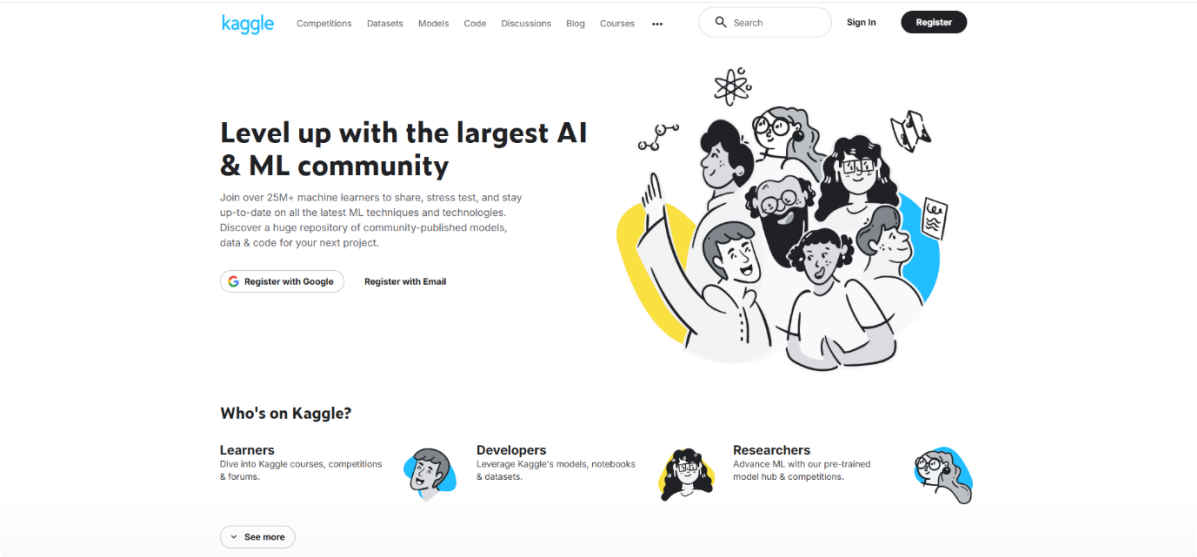

Рисунок 1. Главная страница платформы Kaggle

Другим часто встречающимся в интернете ресурсом, особенно в области **машинного обучения** и интеллектуального анализа данных, является **UCI Machine Learning Repository**. Данный репозиторий поддерживается Калифорнийским университетом в Ирвине. Он включает в себя множество классических датасетов, которые широко используются в научной литературе. Все они тщательно задокументированы и подготовлены для исследовательских целей.

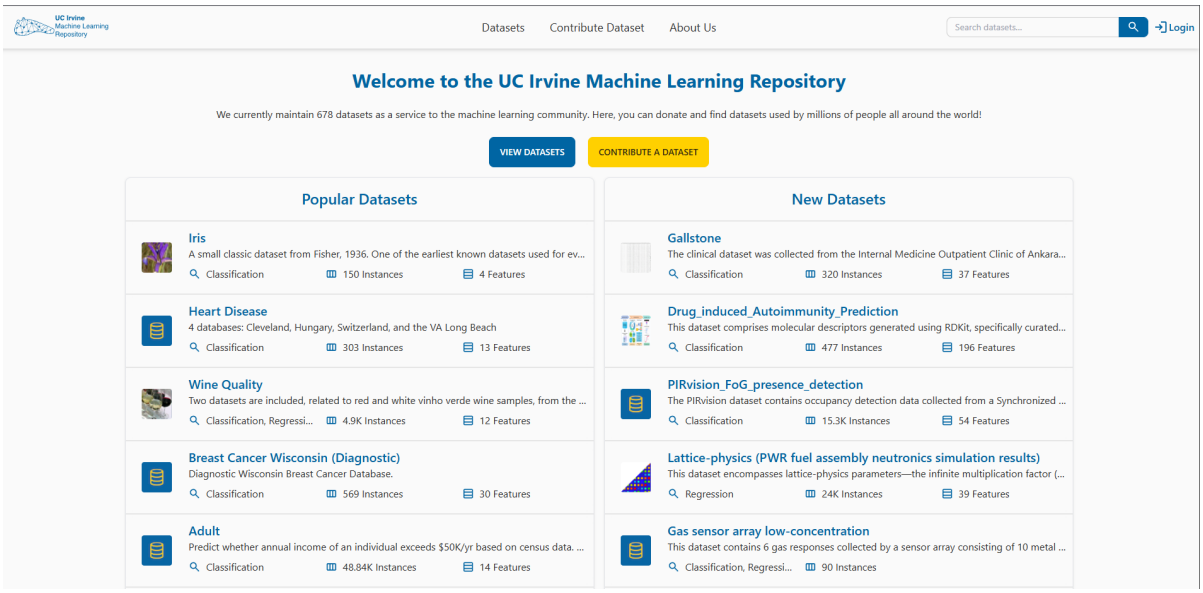

Рисунок 2. Главная страница UCI Machine Learning Repository

Необходимо также упомянуть важный инструмент в поиске подготовленных данных, **Google Dataset Search** – поисковая система от компании **Google**, которая индексирует метаданные датасетов, опубликованных на всевозможных сайтах. Она позволяет находить релевантную информацию, по ключевым словам, независимо от ее первоначального местоположения.

Также стоит затронуть платформу **GitHub**. В первую очередь сервис предназначен для хостинга программного кода и совместной разработки, однако со временем стал популярным местом для размещения различного рода

данных, в том числе и подготовленных датасетов. Здесь встречаются как официальные репозитории исследовательских групп и проектов, так и неформально выложенные наборы данных пользователей. На GitHub можно найти самую разную информацию – от датасетов для обучения нейронных сетей, до результатов парсинга веб-сайтов.

Отдельно можно выделить различного рода **официальные источники [4]**. Исследователи часто прибегают к ним, особенно в области социальных наук, экономики, здравоохранения и демографии, так как, в отличие от пользовательского контента и неформальных исследований, содержат данные, выделяющиеся высокой выборкой, репрезентативностью [5] и соответствием установленным методикам. К ним относятся национальные статистические службы, такие как Росстат в России, Бюро переписи населения США (US Census Bureau), Статистическое управление Великобритании (ONS) или Евростат (Eurostat) в Европейском союзе. Они регулярно публикуют огромные массивы официальных данных по широкому спектру показателей, которые обычно обладают высокой степенью надежности. Кроме вышеперечисленных источников, также можно отметить **Всемирный банк**, который предоставляет обширные статистические данные о развитии стран мира, охватывая экономику, бедность, инфраструктуру, образование и окружающую среду. Сюда также относится Организация Объединенных Наций и ее специализированные учреждения, такие как **Всемирная организация здравоохранения** (ВОЗ), **Продовольственная и сельскохозяйственная организация** (ФАО) или **Детский фонд** (ЮНИСЕФ). Они предлагают уникальные глобальные и региональные статистические сборники по демографии, здоровью населения, продовольственной безопасности и правам человека.

Также исследователей могут заинтересовать корпоративные и отраслевые датасеты от различных компаний, однако тут поиск открытых данных затруднен. Многие технологические компании и медиаорганизации

могут накладывать различные лицензионные ограничения или же делают доступ к данным платным, поэтому далеко не все подобные случаи можно отнести к открытым источникам. Кроме этого, зачастую в таких случаях исследователю предстоит самостоятельно собирать и подготавливать информацию для анализа. Однако можно встретить исключения, как, например, датасет от компании **Uber** – **Uber Movement Data**, содержащий анонимные данные о поездках, что может быть использовано для улучшения городского планирования. Если же какая-нибудь компания решает опубликовать набор данных в открытый доступ для исследовательских целей, то зачастую такие наборы оказываются на вышеупомянутых платформах вроде Kaggle.

Немаловажным источником готовых к анализу датасетов могут служить различные **репозитории, относящиеся к научному сообществу**. Таким может служить крупнейший архив препринтов[1], **arXiv.org**. На сайте часто можно увидеть ссылки на датасеты, использованные авторами в их исследованиях. Также существуют целенаправленно созданные для публикации, сохранения и обмена исследовательскими данными репозитории - **Dryad** и **Figshare**.

Dryad представляет собой курируемый репозиторий, изначально ориентированный на биологические и экологические исследования. Сейчас же на нём можно найти данные из большинства научных областей. Особенностью Dryad является тесная интеграция с научными журналами – многие издательства, например PLOS и eLife, требуют загрузки данных в этот репозиторий перед публикацией статьи. При этом репозиторий использует CC0 (Creative Commons Zero) [6] лицензию. Все наборы данных в Dryad проходят проверку модераторов, за счёт чего обеспечивается соответствие форматов, наличием необходимых метаданных и отсутствием персональных

[1] Препринт — предварительная версия научной статьи, размещаемая в открытом доступе до её рецензирования и публикации в журнале (англ. preprint).

данных. Кроме того, каждый опубликованный датасет имеет свой уникальный идентификатор DOI (Digital Object Identifier).

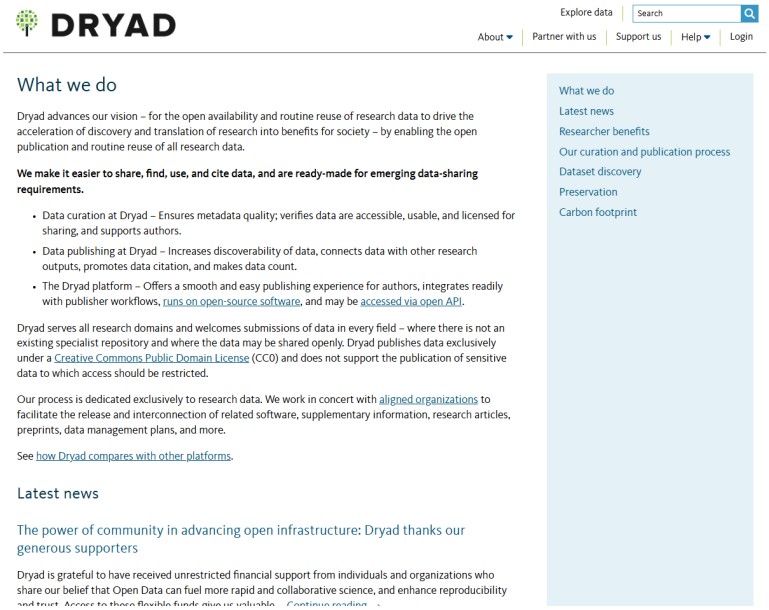

Рисунок 3. Главная страница Dryad

В свою очередь **Figshare** представляет более универсальное решение для различных научных дисциплин. В отличие от Dryad, где используется лицензия CC0, Figshare по умолчанию применяет лицензию CC BY, требующую указания авторства при использовании ресурсов из репозитория в своей работе. Платформа поддерживает хранение данных в самых разных форматах – от традиционных таблиц до трехмерных моделей. Удобной особенностью Figshare является возможность просмотра многих типов данных прямо в браузере, без необходимости скачивать файл.

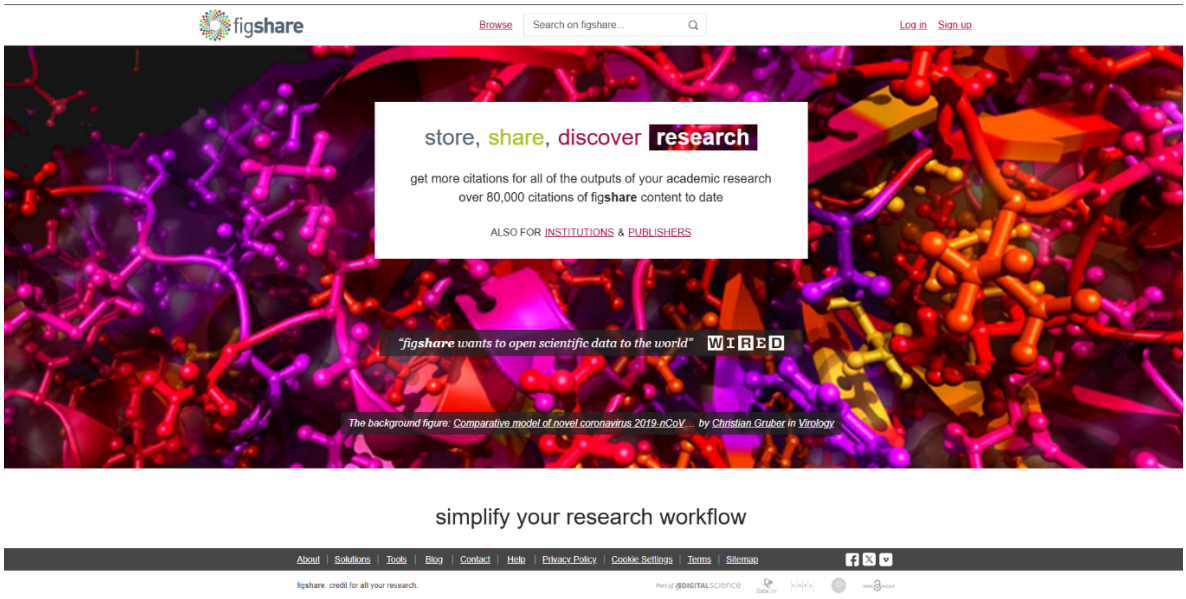

Рисунок 4. Главная страница Figshare

Несмотря на все вышеперечисленные удобства современных сервисов, работа с открытыми источниками сопряжена с радом существенных трудностей. Исследователи обязаны тщательно изучать лицензии, под которыми распространяются данные, чтобы убедиться в законности их использования, особенно в случаях, касающихся публикации и коммерциализации. Существуют глобальные нормы, такие, как например Общий регламент по защите данных (GDPR) (англ. General Data Protection Regulation), действующий в рамках Европейского союза, который накладывает строгие ограничения на работу с персональными данными, что часто делает невозможным использование многих потенциально интересных наборов без специальной обработки. Также острой является проблема качества данных из открытых источников. Они могут содержать наборы с ошибками, пропусками, несогласованными форматами. Однако более важная проблема – сомнительная репрезентативность. Поэтому критическая оценка данных и их валидация становятся обязательным этапом перед использованием датасета в работе.

Сегодня существует огромное разнообразие открытых источников данных. Каждый из них нацелен на свои собственные задачи и имеет свою специфику: от платформы Kaggle и UCI Machine Learning Repository до датасетов государственных организаций, а также научных репозиториев вроде Dryad и Figshare. В условиях роста объема открытых данных важно научиться ориентироваться в этом многообразии. Независимо от выбора источника датасета, стоит внимательно смотреть на его соответствие поставленным задачам.

Список литературы

1. Майер-Шенбергер В., Кукьер К. Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. – Манн, Иванов и Фербер, 2013.
2. Холмс Д. Э. Большие данные: очень краткое введение – Издательство Оксфордского университета, 2017.
3. Абросимова М. А., Власова Л. С. Использование Python при работе с большими данными //Информационные технологии. Проблемы и решения. – 2020. – №. 3. – С. 53-59.
4. Игнатова А. М. Открытые данные как новый способ взаимодействия государства и общества //Исторические, философские, политические и юридические науки, культурология и искусствоведение. Вопросы теории и практики. – 2015. – №. 1-2. – С. 78-80.
5. Нафлик К. Н. Данные: визуализируй, расскажи, используй //КН Нафлик– «Манн, Иванов и Фербер (МИФ). – 2015.
6. Арзуманян А. Б. ЛИЦЕНЗИИ CREATIVE COMMONS: ИСТОРИЯ ПОЯВЛЕНИЯ И ИСПОЛЬЗОВАНИЕ В РОССИИ И ЗА РУБЕЖОМ //Вестник юридического факультета Южного федерального университета. – 2022. – Т. 9. – №. 4. – С. 84-89.

References

1. Mayer-Schoenberger V., Kukier K. Big Data: A revolution that will change the way we live, work and think. – Mann, Ivanov and Ferber, 2013.
2. Holmes D. E. Big Data: a very brief Introduction – Oxford University Press, 2017.
3. Abrosimova M. A., Vlasova L. S. Using Python when working with big data //Information technology. Problems and solutions. - 2020. – No. 3. – pp. 53-59.
4. Ignatova A.M. Open data as a new way of interaction between the state and society //Historical, philosophical, political and legal sciences, cultural studies and art criticism. Questions of theory and practice. - 2015. – № 1-2. – pp. 78-80.

5. Naflik K. N. Data: visualize, tell, use //BOOK Naflik–"Mann, Ivanov and Ferber (MYTH). – 2015.
6. Arzumanyan A. B. CREATIVE COMMONS LICENSES: THE HISTORY OF THEIR APPEARANCE AND USE IN RUSSIA AND ABROAD //Bulletin of the Faculty of Law of the Southern Federal University. – 2022. – Vol. 9. – No. 4. – pp. 84-89.