Публикация наборов данных

А.В. Ермаков [0000-0002-6054-0813]

ИПМ им.М.В.Келдыша РАН
Ermakov@Keldysh.ru

Аннотация. В работе рассмотрены вопросы, связанные с созданием архивов данных, их публикацией и размещением их метаданных в библиографической базе данных Crossref.

Ключевые слова: наборы данных, метаданные публикаций, Crossref, цитирование.

Dataset publication

A.V. Ermakov [0000-0002-6054-0813]

Keldysh Institute of Applied Mathematics RAS
Ermakov@Keldysh.ru

Abstract. The paper deals with issues related to the creation of data archives, their publication and placement of their metadata in the Crossref bibliographic database.

Keywords: dataset publication, Crossref, citations.

Публикация научных материалов имеет большую историю. Развитию направления научной литературы способствовало появление научных журналов, публикация монографий и научные конференции, материалы которых сначала носили устный характер, а затем также стали размещаться на «бумажных носителях» в виде сборников статей.

С точки зрения популяризации наборы научных данных значительно отстали от научных статей. Если текстовые табличные данные (например, таблица умножения, таблицы Брадиса) или графики еще можно представить на бумажных носителях, то более сложные многомерные массивы данных или двоичные данные публиковать сложнее. Их и размещать-то на каких-либо носителях до появления вычислительной техники было практически невозможно.

Появление компьютеров и Интернета решило проблему размещения цифровой информации и многомерных наборов данных, количество и

объем которых с развитием мощных вычислительных систем значительно выросли. Для работы с полученными научными результатами появились удобные пользовательские инструменты, в том числе графические системы.

Но работать с научными результатами такого рода могла только небольшая группа научных сотрудников, связанных с теми, кто получил и сохранил результаты вычислений, так как необходимо знать расположение наборов данных, их формат и многие другие особенности.

Появление баз данных и репозиториев для хранения данных позволило более широкому кругу ученых получить возможность исследовать накопленные результаты. Это дало мощный толчок работам в различных научных направлениях (астрономии, медицине, социологии и др.) как за рубежом, так и в нашей стране.

Однако вопросы, связанные с публикацией научных данных, по-прежнему остаются во многом нерешенными. Для научной статьи формат публикации, библиографическая ссылка и т.п. определяются требованиями издания, национальными и международными стандартами. Для наборов данных все это находится в зачаточном состоянии. Хотя результаты вычислительных экспериментов являются первичными в смысле выполнения работ по договорам и грантам, сослаться на них в отчетных материалах напрямую пока нельзя. Представить полученные результаты можно только в научной статье, опубликованной в журнале или сборнике материалов конференции. Но публикация научной статьи требует время и на ее подготовку, и на размещение в каком-либо научном издании. При этом в значительной степени теряется актуальность.

Оперативное размещение полученных научных результатов в архиве данных, краткая аннотация, описание формата или указание инструмента, дающего исследователю доступ к указанным наборам данных, в значительной степени решают вопросы ускорения и упрощения доступа к информации.

Однако в нашей стране создание архивов данных и инструментов работы с ними находится пока на нуле. То есть использовать чужие результаты в своих исследованиях — такие возможности есть, и примеры найти несложно. А вот размещение своих результатов в зарубежных архивах вызывает много вопросов и по поводу авторства, и по поводу разработки и использования собственных инструментов работы с данными.

В данной статье на примерах мы расскажем о том, как и какие архивы создаются в мире, какие есть инструменты работы с данными и о нашем эксперименте по созданию архива и публикации наборов данных.

Репозитории

Репозитории представляют собой хранилища данных различной природы, предназначенные для размещения наборов данных и последующего их использования возможно с помощью некоторых специализированных инструментов.

Появление облачных вычислений привело к появлению систем хранения данных, помимо традиционной локальной инфраструктуры, размещающихся в различных местах, включая локальные решения, частные и общедоступные облака.

Современные хранилища данных предназначены для обработки структурированных и неструктурированных данных, таких как видео, файлы изображений, данные с различных приборов, датчиков, а главное – это результаты компьютерных расчетов. Без хранилища данных очень сложно объединять данные из неоднородных источников, обеспечивать нужный формат для последующей обработки, получать актуальное и долгосрочное представление о данных во времени, объединять и анализировать результаты нескольких взаимосвязанных вычислительных экспериментов.

Хранилище данных обеспечивает множество преимуществ. Вот некоторые из них:

- Повышение качества научных исследований. При использовании хранилищ данных любой научный сотрудник соответствующей квалификации может проверить корректность выводов, представленных в научной публикации, опирающейся на указанный набор данных. Более того, заинтересованный исследователь может дополнить и расширить исследования в данном направлении, имея доступ ко всей исходной информации.
- Ускорение выполнения запросов. Хранилища данных создаются специально для быстрого извлечения и анализа данных. При использовании хранилищ можно очень быстро запрашивать большие объемы консолидированных данных.
- Повышение качества данных. Перед загрузкой в хранилище для удобства последующей обработки рекомендуется обеспечить преобразование данных в согласованный формат, упрощающий для широкой научной общественности доступ к представленным материалам.
- Ретроспективный анализ. Хранилище содержит большие объемы ранее полученных данных и позволяет исследователям изучать прошлые тенденции и проблемы, делать прогнозы и постоянно совершенствовать инструменты для наблюдений.

Типичное хранилище данных состоит из четырех основных компонентов: центральной базы данных, инструментов ETL (извлечение, преобразование, загрузка), метаданных и инструментов доступа. Все эти

компоненты разработаны с прицелом на обеспечение максимальной скорости, что позволяет быстро получать результаты и оперативно анализировать данные.

Метаданные — это данные о данных. Они определяют источник, механизм использования, значения и другие функции наборов данных в хранилище данных. Существуют библиографические метаданные, описывающие название, авторов, дату публикации и т.п. и технические метаданные, которые указывают способ доступа к данным, включая их местоположение и структуру.

Примеры архивов для хранения наборов данных

Рассмотрим правила и особенности некоторых популярных архивов.

Zenodo (www.zenodo.org) — это универсальный репозиторий с открытым доступом, разработанный в рамках европейской программы OpenAIRE и управляемый ЦЕРН. Основными поставщиками данных этого архива являются Управления научных миссий НАСА и Сообщество исследователей коронавирусных заболеваний - COVID-19.

Политика архива позволяет исследователям депонировать результаты исследований, наборы данных, исследовательское программное обеспечение, отчеты и любые другие цифровые материалы, связанные с исследованиями. Для каждого представления создается постоянный идентификатор цифрового объекта (DOI), что позволяет легко цитировать хранимые элементы.

Однако есть некоторые существенные ограничения, которые пользователи должны учитывать в своей работе (https://help.zenodo.org/):

- После того, как запись опубликована, автор больше не может изменять файлы в записи.
- В исключительных случаях разрешается вносить небольшие изменения в файлы записи, если запись была опубликована недавно (менее одной недели). Если автор заметил ошибки, такие как опечатки, случайное упущение важных файлов/включение скрытых файлов или конфиденциальных файлов, и хотел бы исправить их, следует связаться с администрацией.
- Размещение в архиве материалов с DOI, отличных от Zenodo, невозможно.

Figshare (https://figshare.com) – репозиторий, разработанный и поддерживаемый в рамках проекта Digital Science, направленного на развитие исследований экосистем.

Figshare — это научный репозиторий, где пользователи могут сделать доступными результаты своих исследований, могут сохранять данные и обмениваться результатами своих исследований, включая

рисунки, базы данных, изображения и видео. Загруженным научным материалам автоматически присваиваются DOI.

Хорошими примерами размещенных в данном архиве наборов данных являются показатели качества статей в Википедии (WikiRank https://figshare.com/authors/Wiki_Rank/5909681).

Digitalrocksportal - DRP (https://www.digitalrocksportal.org)

Digital Rocks — это портал данных для долговременного хранения, поиска, обмена, организации и анализа изображений различных пористых микроструктур. Он направлен на расширение исследовательских ресурсов для моделирования / прогнозирования свойств пористых материалов в области нефтяной, гражданской и экологической инженерии, а также геологии.

Портал Digital Rocks был разработан и финансируется в настоящее время за счет грантов Национального научного фонда США (NSF).

Эта платформа позволяет управлять, сохранять, визуализировать и выполнять базовый анализ имеющихся изображений пористых материалов и проведенных с ними экспериментов, а также любые сопутствующие измерения (пористость, капиллярное давление, проницаемость, электрические и упругие свойства и т. д.), необходимые для валидации подходов к моделированию, а также масштабирование и построение более крупных (гидро)геологических моделей.

В DRP пользователи могут организовать свои данные в соответствии со структурой своего исследовательского проекта, чтобы различать: исходные данные и данные для анализа, полученные в результате экспериментов и/или моделирования.

У всех авторов и участников проектов есть возможность вводить данные и метаданные в любом порядке в течение жизненного цикла своего исследовательского процесса, просматривать и создавать отношения между данными и метаданными.

Для задания метаданных наборов данных в DRP используется схема компании DataCite, которая является для портала провайдером DOI.

На портале пользователи могут хранить, систематизировать и описывать свои данные конфиденциально, пока они не будут готовы к публикации. Это позволяет сохранять и накапливать данные по мере продвижения исследований, а также редактировать метаданные или добавлять документацию своего исследовательского проекта.

WDC-climate (https://www.wdc-climate.de) – репозиторий, созданный Немецким центром климатических вычислений ((DKRZ: Deutsches Klimarechenzentrum GmbH) для архивирования больших наборов исследовательских данных, которые имеют отношение к изучению

климата в рамках Мирового центра данных по климату (World Data Centre for Climate –WDCC).

Базовый набор метаданных, предоставляемый производителем данных, описывает данные и позволяет легко найти их в архиве, все метаданные находятся в открытом доступе.

Архив дает возможность долгосрочного архивирования для данных проекта DKRZ HPC (высокопроизводительные вычисления), а также для данных из внешних источников.

Архив предоставляет широкий спектр возможностей по поиску данных, отсортированных по проектам, темам, временному и пространственному охвату или другим аспектам.

Процесс публикации данных аналогичен публикации статьи в научном журнале. Данные должны соответствовать определенным требованиям качества. Постоянный доступ к данным предоставляется присвоенным уникальным идентификатором DOI — данные и метаданные остаются неизменными.

Как показано в материалах представленных нами примерах архивов наборов данных, администрация репозиториев предоставляет пользователям некоторый набор инструментов для размещения, хранения, поиска и публикации данных. Всем наборам присваивается цифровой идентификатор DOI, во многих случаях изменение данных и метаданных после публикации не допускается.

Правила и особенности размещения наборов данных в архивах

Любой набор данных, размещаемый в архиве, имеет название, авторов, дату публикации, краткую аннотацию, ключевые слова, а также лицензионные ограничения на использование данных материалов.

Размещенному в архиве набору данных автоматически присваивается цифровой идентификатор (DOI). Более того, если предполагается научная публикация, опирающиеся на материалы архивируемого набора данных, то в большинстве случаев можно зарезервировать DOI до непосредственной загрузки материалов.

Для каждого опубликованного набора данных система ведет учет количества скачиваний и просмотров. Однако допустима работа и с закрытыми от общего доступа архивами.

В качестве примера оформления опубликованного в архиве набора данных используем материалы одного из ранее рассмотренных репозиториев (Рис. 1)

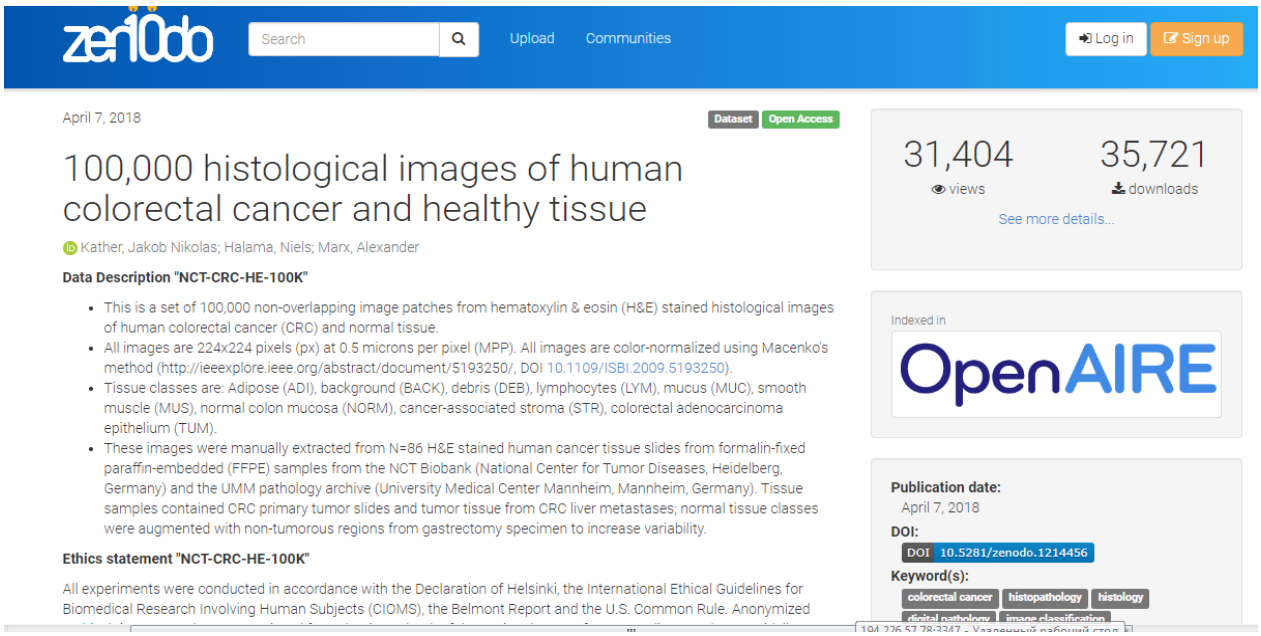

Рис. 1 Пример оформления размещенных в архиве наборов данных.

Архивы данных в ИПМ

В качестве эксперимента мы создали в Институте архив данных, связанный расчетами термодинамического равновесия никелевого кластера (https://polyakov.imamod.ru/arc/data/index.html). Разместили там два набора данных и присвоили им DOI. При этом нам удалось разобраться с атрибутами метаданных, характерными именно для регистрации наборов данных.

<database> является контейнером для всей информации о «группе» наборов данных. База данных верхнего уровня может быть функциональной базой данных или абстракцией, действующей как коллекция (так же, как журнал — это коллекция статей).

Отдельные записи набора данных фиксируются внутри <dataset> элемента.

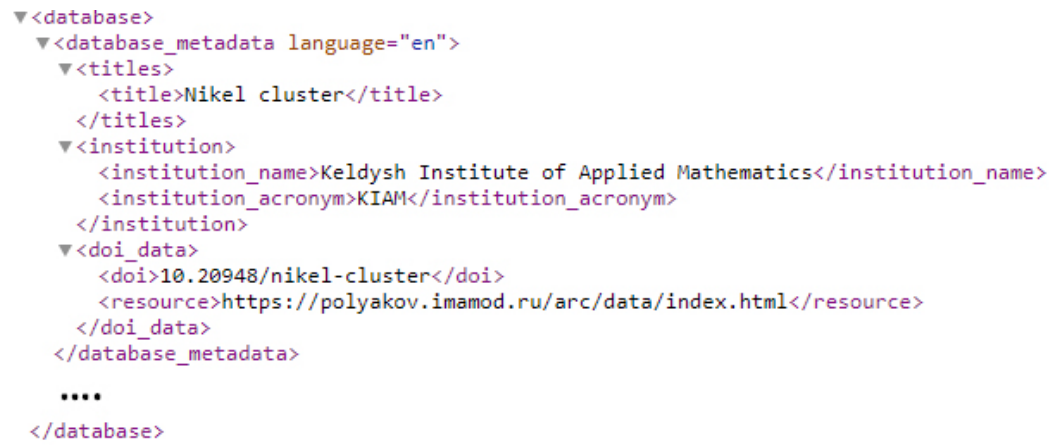

Рис. 2. Пример оформления размещенных в архиве наборов данных.

Т.е. в примере с нашим архивом:
10.20948/nikel-cluster – это database (или коллекция);
10.20948/nikel-cluster-CUBE2023 – это 1-й dataset этой коллекции;
10.20948/nikel-cluster-BALL2023 – это 2-й dataset этой коллекции.

Отметим, что также, как в научных статьях, в наборах данных используется цитирование – как в статье можно сослаться на набор данных, так и в описании набора данных можно сослаться на статью, в которой этот набор описывается или используется.

Кто уже цитирует данные

В заключении приведем Top10 издательств, в которых авторы статей наиболее активно обращаются (цитируют) к наборам данных, размещенным в различных репозиториях (с разбивкой по префиксу DOI, поэтому некоторые издатели перечислены дважды).

Префикс DOI	Издательство	Количество
10.1038	Springer Science and Business Media LLC	7174
10.1016	Elsevier BV	6527
10.1007	Springer Science and Business Media LLC	4748
10.5194	Copernicus GmbH	3017
10.1080	Informa UK	2346
10.1177	SAGE Publications	2082
10.1002	Wiley	2048
10.1111	Wiley	1888
10.1108	Emerald	1876
10.3390	MDPI AG	1827

Рис.3. Топ-10 участников, депонирующих цитирование данных за ноябрь-май 2022 г.

В научных материалах, публикуемых в нашем Институте, мы нашли 5 публикаций, цитирующих наборы данных, – 2 ссылки в сборниках нашей конференции и 3 статьи в сборниках Графикона.

Литература

References