



ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 62 за 2023 г.



ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

А.В. Ермаков

Публикация и цитирование наборов данных

Статья доступна по лицензии
[Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)



Рекомендуемая форма библиографической ссылки: Ермаков А.В. Публикация и цитирование наборов данных // Препринты ИПМ им. М.В.Келдыша. 2023. № 62. 16 с.
<https://doi.org/10.20948/prepr-2023-62>
<https://library.keldysh.ru/preprint.asp?id=2023-62>

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.В. Ермаков

**Публикация и цитирование
наборов данных**

Москва — 2023

Ермаков А.В.

Публикация и цитирование наборов данных

В работе рассмотрены вопросы, связанные с созданием архивов данных, их публикацией, размещением метаданных в библиографической базе данных Crossref и цитированием наборов данных в научных публикациях. Приведены примеры популярных мировых архивов данных и рассказано об эксперименте по созданию и публикации наборов данных в ИПМ.

Ключевые слова: публикация наборов данных, метаданные публикаций, Crossref, цитирование наборов данных.

Alexey Viktorovich Ermakov.

Publishing and citing datasets

The paper addresses issues related to the creation of data archives, their publication, placement of metadata in the Crossref bibliographic database and citation of datasets in scientific publications. Examples of popular world data archives are given and an experiment on creating and publishing datasets in KIAM is described.

Key words: publishing datasets, publication metadata, Crossref, citing datasets.

Оглавление

Введение	3
1. Репозитории	4
2. Примеры архивов для хранения наборов данных	5
3. Архивы данных в ИПМ	10
4. Цитирование наборов данных	12
5. Индекс цитирования научных данных.....	13
6. Кто уже цитирует данные	15
Список литературы.....	16

Введение

Публикация научных материалов имеет большую историю. Развитию направления научной литературы способствовали 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 репозиторий 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 времени 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.

Автор выражает глубокую благодарность д.ф.-м.н. С.В. Полякову и д.ф.-м.н. В.О. Подрыге за большую помощь, консультации и замечания при подготовке данного материала.

1. 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 (Extract, Transform, Load — то есть инструментов для извлечения, преобразования и загрузки), 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.

2. Примеры архивов для хранения наборов данных

Рассмотрим правила и особенности некоторых популярных архивов.

Zenodo (<https://zenodo.org>) — это универсальный репозиторий с открытым доступом, разработанный в рамках европейской программы OpenAIRE и управляемый ЦЕРН. Основными поставщиками данных этого архива являются Управление научных миссий NASA 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.

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

Прежде всего отметим типичные библиографические атрибуты — название, авторы, краткая аннотация.

Для каждого опубликованного набора данных система ведет учет количества скачиваний и просмотров (однако допустима работа и с закрытыми

от общего доступа архивами). И на рис. 1 легко отметить достаточно высокий интерес к данному набору данных.

В нижней части рисунка уже привычные библиографические параметры — дата публикации, DOI и ключевые слова.

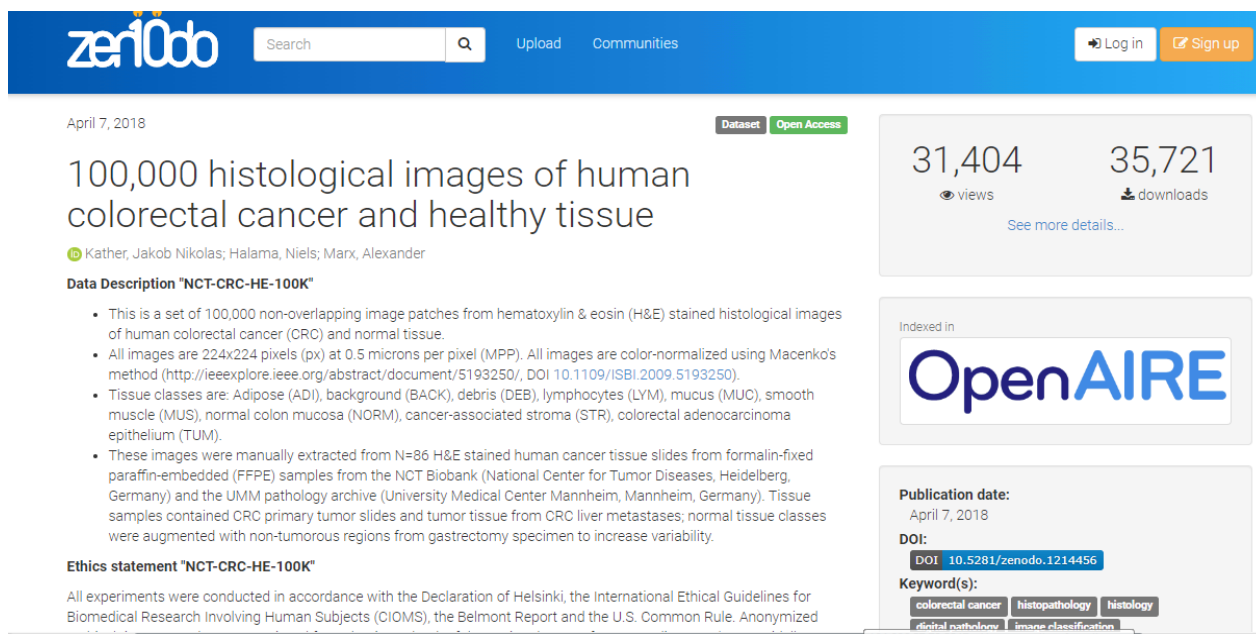


Рис. 1. Пример оформления размещенных в архиве Zenodo наборов данных.

Однако есть некоторые существенные ограничения в политике данного архива, 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 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.

Типичным примером может служить набор данных, представленный на рис. 2. Здесь так же, как и в архиве Zenodo (и практически во всех других научных архивах), представлены ключевые атрибуты публикации: авторы, название, краткая аннотация представленного материала. Также присутствует и статистическая информация, указывающая на высокий интерес (в данном конкретном случае) читателей, и число цитирований (которых на данный момент нет).



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

Но, в отличие от политики Zenodo, в архиве Figshare допускается перезапись сохраняемых данных с указанием номера версии и даты (все предыдущие версии также доступны).

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

Digital rocks portal – DRP (<https://www.digitalrocksporal.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.

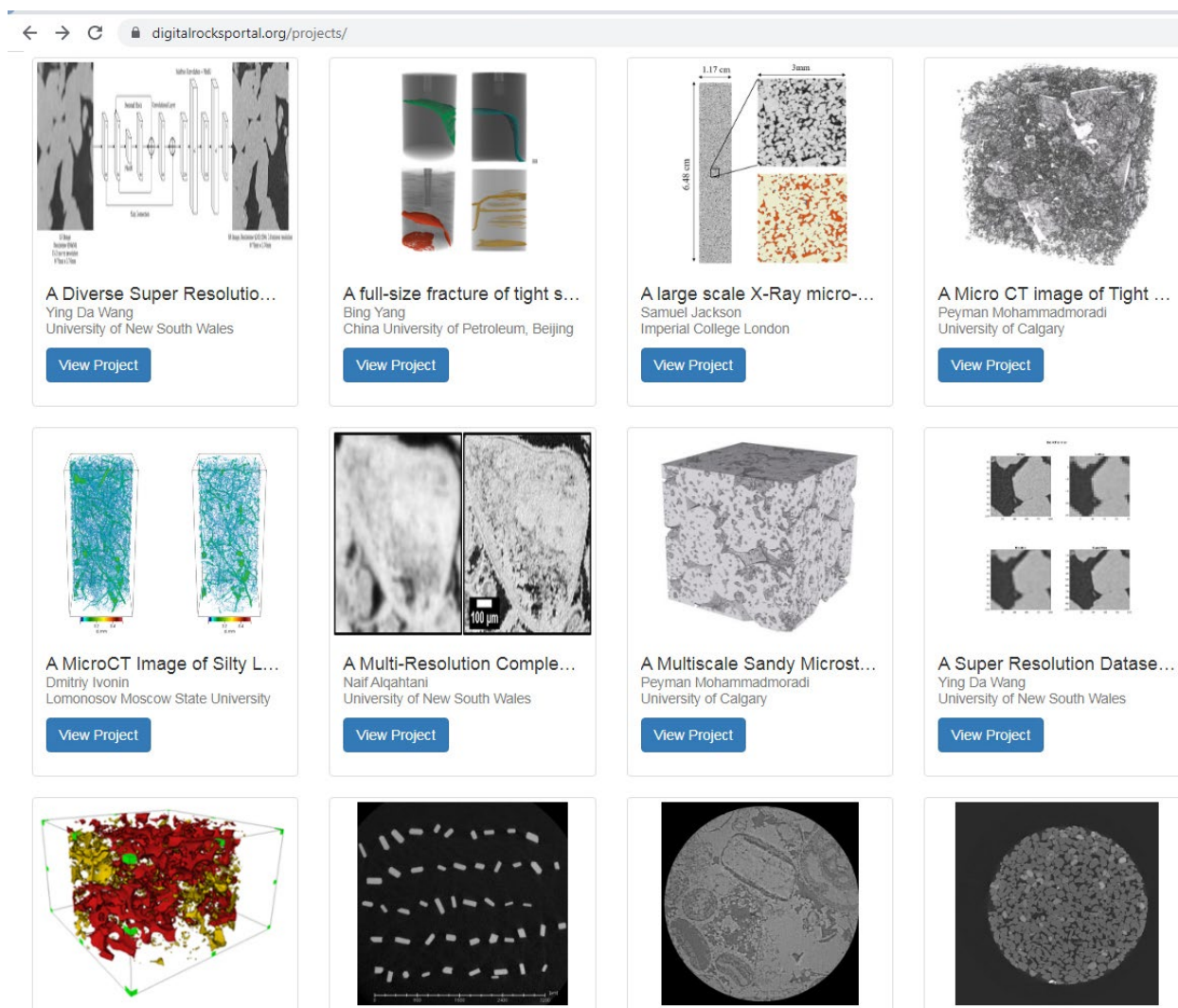


Рис. 3. Пример "иллюстрированного оглавления" архива Digital Rocks.

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 выполнять 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 их экспериментов и/или моделирования. То есть "проект" является в данном архиве ключевым элементом архитектуры (рис. 4).

The screenshot shows the Digital Rocks Portal interface. At the top is a navigation bar with the logo, a search bar, and links for 'Browse Projects', 'How to Use', 'About', 'Log in', and 'Register'. The main title of the project is 'Bidisperse sphere packs generated under gravity'. Below the title are two tabs: 'Project' (selected) and 'Related Publications'. The 'Project' tab contains a 3D visualization of red spheres of different sizes on a green background. To the right of the visualization is a 'Description' section with text about the generation of bidisperse sphere packs. Further right is a metadata section with fields for 'Author' (Abhishek Bihani), 'Collaborators' (Hugh Daigle), 'Created' (March 5, 2019), 'License' (ODC-BY 1.0), and 'Digital Object Identifier' (10.17612/P74T20). Below the description is a 'Datasets' section with a table listing eight data samples, each with an 'Origin Data' link. To the right of the datasets is a 'Download Project' section with a description of the archive and a 'Download' button.

Bidisperse sphere packs generated under gravity

Description

Bidisperse sphere packs consisting of small (clay) and large (silt) grains were generated under the effect of gravity to model mudrocks and study the petrophysical properties. Multiple, periodic, dense, bidisperse sphere packs were created with large grain volume concentrations (VL) increasing from 0 to 100 %. This was done for large to small grain radius ratios of 2:1, 5:1, 7:1, and 10:1 with constant sphere pack length 10 units in x, y, z directions. The data for each sphere pack consists of the x, y, z coordinates of the sphere centers and the sphere radii stored in .csv files. Cross-sectional images visualized in ParaView are shown for clarity.

Datasets

Sample	2:1 radius ratio sphere packs
Origin Data	2to1_VL1
Origin Data	2to1_VL10
Origin Data	2to1_VL20
Origin Data	2to1_VL30
Origin Data	2to1_VL40
Origin Data	2to1_VL50
Origin Data	2to1_VL60
Origin Data	2to1_VL70

Author
Abhishek Bihani (University of Texas at Austin)

Collaborators
Hugh Daigle (The University of Texas at Austin)

Created
March 5, 2019

License
ODC-BY 1.0

Digital Object Identifier
10.17612/P74T20

[Data Citation](#)

Download Project

The downloadable archive contains all project data; the size of the archive file for this project is **0.30 GB**.

[Download](#)

Рис. 4. Пример оформления размещенных в архиве Digital Rocks наборов данных.

Лучше всего, если человек, который является автором данных (справа на рис. 4) и будет первым автором в цитировании данных после публикации проекта, одновременно является владельцем проекта в DRP. Владелец проекта может добавлять соавторов, которые смогут редактировать проект и добавлять данные. Владелец данных (скажем, главным исследователем более крупной исследовательской группы) не всегда является человек, загружающий файл. В этом случае владелец данных должен зарегистрироваться, создать проект под желаемым именем и добавить загрузчика данных (который также уже зарегистрировался) в качестве соавтора. После этого загрузчик данных может добавлять и систематизировать данные. Если автор проекта доволен макетом данных проекта, то он может запросить публикацию (присвоить DOI). Далее можно добавлять соавторов (Collaborators, рис. 4), независимо от того, являются ли они зарегистрированными пользователями в DRP или нет, чтобы они отображались в будущих цитированиях данных.

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 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 НРС (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 примеров архивов наборов данных, администрация репозитория 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.

3. Архивы данных в ИПМ

В качестве эксперимента мы создали в Институте архив данных, связанный с 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.

Например, атрибут `<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.

```
▼<database>
  ▼<database_metadata language="en">
    ▼<titles>
      <title>Nickel cluster</title>
    </titles>
    ▼<institution>
      <institution_name>Keldysh Institute of Applied Mathematics</institution_name>
      <institution_acronym>KIAM</institution_acronym>
    </institution>
    ▼<doi_data>
      <doi>10.20948/nickel-cluster</doi>
      <resource>https://polyakov.imamod.ru/arc/data/index.html</resource>
    </doi_data>
  </database_metadata>
  ....
</database>
```

Рис. 5. Пример оформления размещенных в архиве наборов данных.

Т.е. в примере с нашим архивом:

10.20948/nickel-cluster— это database (или коллекция)[1];

10.20948/nickel-cluster-CUBE2023 — это 1-й dataset этой коллекции [2];

10.20948/nickel-cluster-BALL2023 —это 2-й dataset этой коллекции [3].

Отметим, что, так же как в научных статьях, в наборах данных может использоваться 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.

```
▼<database>
  ....
  ▼<dataset dataset_type="collection">
    ....
    ▼<citation_list>
      ▼<citation key="1">
        ▼<unstructured_citation>
          Подрыга В.О., Поляков С.В. Молекулярно-динамическое моделирование
          процесса установления термодинамического равновесия нагретого
          Препринты ИПМ им. М.В.Келдыша. 2014. № 41. 20 с.
          https://library.keldysh.ru/preprint.asp?id=2014-41
        </unstructured_citation>
      </citation>
    </citation_list>
  </dataset>
</database>
```

Рис. 6. Пример оформления цитирования в метаданных.

4. Цитирование наборов данных

Цитирование наборов данных позволяет исследователям ссылаться на данные как на отдельные исследовательские продукты, получать признание за архивирование и обмен данными, а также отслеживать использование и влияние своих данных.

Однако, в отличие от цитирования научных статей, стандарты цитирования данных пока проработаны, на наш взгляд, недостаточно. Основная рекомендация в этом случае – использование форматов цитирования, рекомендованных издательством [4].

Таким образом, предлагаемый формат для ссылки на набор данных может выглядеть следующим образом:

- Год (или диапазон лет, где это применимо);
- Название набора данных;
- Версия (в круглых скобках, где это необходимо);
- Формат [в квадратных скобках];
- Издательство;
- DOI или URL-адрес.

Ниже приведен пример ссылки на набор данных, оформленной в соответствии с вышеуказанными рекомендациями.

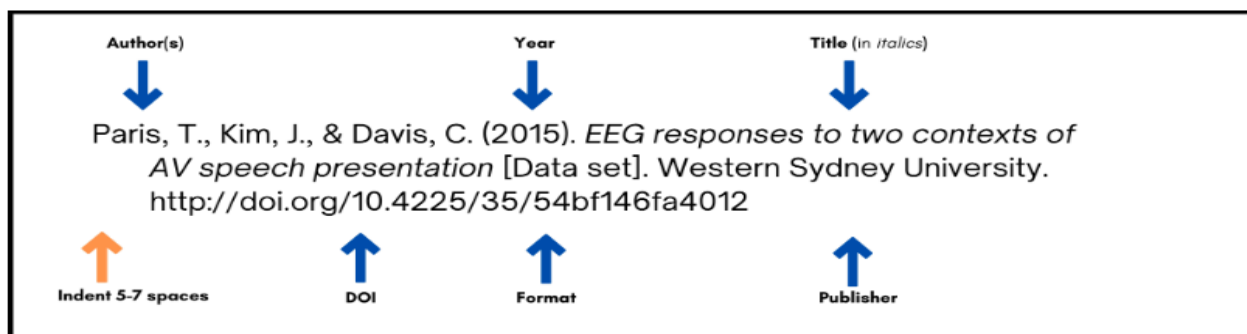


Рис. 7. Пример оформления ссылки на набор данных.

Тогда ссылки на наборы данных, описанные в предыдущем разделе, будут выглядеть следующим образом:

1. С.В. Поляков, В.О. Подрыга. (2023). Расчетные данные по никелевому кластеру кубической формы. [Набор данных]. Институт прикладной математики им. М.В. Келдыша.

<https://doi.org/10.20948/nickel-cluster-CUBE2023>

2. С.В. Поляков, В.О. Подрыга. (2023). Расчетные данные по никелевому кластеру сферической формы. [Набор данных]. Институт прикладной математики им. М.В. Келдыша.

<https://doi.org/10.20948/nickel-cluster-BALL2023>

Но большинство поисковых систем «не заточено» на формат ссылки на набор данных, размещенных в хранилищах. Например, Google в ответ на запрос, содержащий авторов и название набора данных, выдаст список публикаций, ссылающихся на эти данные. А непосредственно репозиторий и локацию в нем, возможно, и не найдет. Это потому, что контекст запроса, скорее всего, будет частично совпадать с библиографическими ссылками в найденных статьях. А правильно воспринимать метаданные хранящихся наборов данных поисковые системы пока умеют (на наш взгляд) недостаточно хорошо.

Конечно, положение спасает правильно оформленная библиографическая ссылка в найденных статьях – имея DOI и/или URL, найти искомые данные не составит труда.

5. Индекс цитирования научных данных

Как мы уже отмечали, данные предоставляют более полную картину исследований, проводимых в рамках исследуемой темы и описанных в научных статьях. Во многом благодаря правительственным предписаниям по обеспечению свободного доступа к данным исследований во всем мире растет число хранилищ данных, которые пополняются результатами фото-видео-аудио-наблюдений, компьютерных расчетов, медицинских процедур и т.д.

Ссылки между наборами данных и опубликованными исследованиями позволяют находить соответствующие данные, в то время как те, кто собирает и размещает данные, исследуя показатели статистики обращений к архивам данных, могут оценивать востребованность своей работы.

Индекс цитирования данных (Data Citation Index, DCI [5]) обеспечивает более простой доступ к данным исследований из различных международных хранилищ по всем дисциплинам на фоне постоянного роста объема цифровых веб-ресурсов.

Политики оценки и выбора репозиторий для Индекса цитирования данных разработаны и запущены в 2012 году в компании Clarivate Analytics Web of Science [6].

Авторы указанного проекта под исследованием данных понимают изучение результатов экспериментов либо фото-видео-аудио-материалов, размещенных в репозиториях, с соответствующими метаданными и/или программным обеспечением, которое может быть использовано при анализе данных. Исследования данных могут быть цитируемым объектом в литературе и могут содержать цитируемые ссылки, прикрепленные к их метаданным, а также информацию о таких аспектах, как основные исследователи, информация о финансировании, предметные термины, географический охват и т. д. Уровень предоставляемых метаданных варьируется в зависимости от политики администрации хранилища.

Процесс оценки потенциальных участников Индекса цитирования данных осуществляется постоянно, при этом новые репозитории добавляются в DCI еженедельно [7]. Более того, список архивов, уже вошедших в Индекс цитирования данных, постоянно пересматривается. Проиндексированные хранилища контролируются, чтобы гарантировать, что они остаются доступными и соответствуют требуемым стандартам.

DCI включает данные исследований из трех основных предметных областей: наука и технологии, социальные науки и искусство и гуманитарные науки. Отдельные репозитории могут быть мультидисциплинарными, междисциплинарными или могут иметь узкую направленность. Имея в своем распоряжении огромный объем данных, редакторы Clarivate Analytics имеют все возможности для выявления новых тем и актуальных научных направлений.

При оценке архивов данных на предмет включения в DCI учитывается множество факторов, как качественных, так и количественных: основные издательские стандарты репозитория, его редакционная политика, международное разнообразие его авторства и связанные с ним данные цитирования. Ни один фактор не рассматривается изолированно, но путем объединения и взаимосвязи данных могут быть определены общие сильные и слабые стороны репозитория.

Надежность сохранения данных репозитория является одним из основных критериев в процессе оценки. Репозиторий должен продемонстрировать долговечность, чтобы его можно было первоначально включить в Индекс цитирования данных. Clarivate Analytics также проверяет, идет ли поступление новых данных – постоянный поток вновь депонированных данных считается показателем того, что ресурс в данный момент активен. Как правило, данные следует хранить в репозитории, а не в хранилище, просто хранящем метаданные и предоставляющем веб-ссылку на удаленный/внешний источник данных. Это обеспечивает надежное цитирование данных, что позволяет использовать метрики цитирования и повторное использование данных. В идеале должно быть указано четкое определение процесса публикации данных с указанием принадлежности поставщика/создателя данных. Если репозиторий выбран для включения в Индекс цитирования, то все депонированные данные будут включены в индекс цитирования данных.

Как уже отмечалось, Индекс цитирования данных призван способствовать цитированию данных и связыванию данных с исследовательской литературой. С этой целью особое внимание уделяется хранилищам, которые показывают происхождение информационных материалов и сопровождаются указаниями о финансировании грантов. Эксперты Clarivate Analytics в первую очередь фокусируют свое внимание на репозиториях, которые публикуют метаданные на английском языке или, по крайней мере, позволяют предоставлять достаточную описательную информацию (метаданные) на английском языке.

Хотя рецензирование депонированных данных ни в коем случае не является обязательным, применение процесса рецензирования при публикации данных является еще одним показателем высоких стандартов хранилища, означает общее качество представленных данных и полноту любых цитируемых ссылок.

6. Кто уже цитирует данные

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

Таб. 1. Топ-10 участников, депонирующих цитирование данных, за ноябрь 2022 г. – май 2023 г.

Префикс DOI	Издательство	Количество
10.1038	SpringerScienceandBusinessMediaLLC	7174
10.1016	ElsevierBV	6527
10.1007	SpringerScienceandBusinessMediaLLC	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

Бросается в глаза то, что общее число цитирований исследуемых наборов данных в журналах ведущих мировых издательств значительно уступает вниманию (количеству просмотров и количеству скачиваний — рис. 1 и рис. 2) к архивам данных, на которые опираются опубликованные исследования. На основе анализа значительного количества публикаций, описывающих исследования, опирающиеся на наборы данных, можно сделать следующие выводы:

- Чаще всего авторы, подробно описывая исследуемый набор данных, вообще не ссылаются на него.
- Следующая по распространенности ошибка — цитирование не самого набора данных, а другой публикации, описывающей его. При этом авторы такой публикации могут как совпадать с авторами набора данных, так и отличаться.
- Третья по распространенности проблема связана с тем, что исследования проводятся с данными, просто размещенными в интернете. В этом случае даже указанный URL набора данных в библиографическом списке не может считаться корректным цитированием, так как такого рода набор

данных не считается корректно опубликованным (не указаны метаданные публикации, не присвоен DOI).

И в заключение следует сделать вывод, что необходимо продолжать изучение и обсуждение различных вопросов, связанных с публикацией и цитированием наборов данных.

Список литературы

1. С.В. Поляков, В.О. Подрыга. (2023). Наборы расчетных данных по никелевому кластеру. [Набор данных]. Институт прикладной математики им. М.В. Келдыша.
<https://doi.org/10.20948/nickel-cluster>
2. С.В. Поляков, В.О. Подрыга. (2023). Расчетные данные по никелевому кластеру кубической формы. [Набор данных]. Институт прикладной математики им. М.В. Келдыша.
<https://doi.org/10.20948/nickel-cluster-CUBE2023>
3. С.В. Поляков, В.О. Подрыга. (2023). Расчетные данные по никелевому кластеру сферической формы. [Набор данных]. Институт прикладной математики им. М.В. Келдыша.
<https://doi.org/10.20948/nickel-cluster-BALL2023>
4. Cousijn, H., Kenall, A., Ganley, E. et al. A data citation roadmap for scientific publishers. *Sci Data* 5, 180259 (2018).
<https://doi.org/10.1038/sdata.2018.259>
5. Индекс цитирования данных (Data Citation Index)
<https://clarivate.com/cis/solutions/data-citation-index/>
6. Clarivate Analytics Web of Science
<https://clarivate.com/webofsciencegroup/essays/the-repository-selection-process/>
7. Список архивов наборов данных, вошедших в DCI.
<https://clarivate.com/webofsciencegroup/master-data-repository-list/>