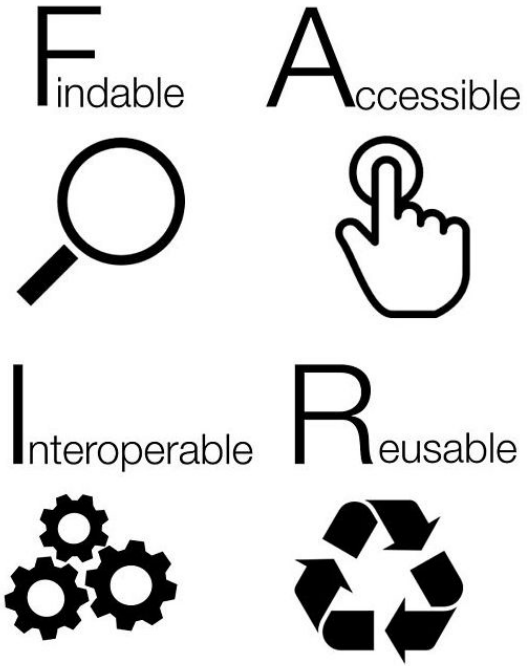
Биоәртүрлілікті зерттеудегі цифрлық технологиялар
Digital technologies in biodiversity research
Цифровые технологии в исследовании биоразнообразия

ЛЕКЦИЯ 5

FAIR концепция данных и биоразнообразие



Слайды CC BY:
Dag Endresen, GBIF Norway
Наталья Иванова,
Максим Шашков

План лекции

Открытая наука

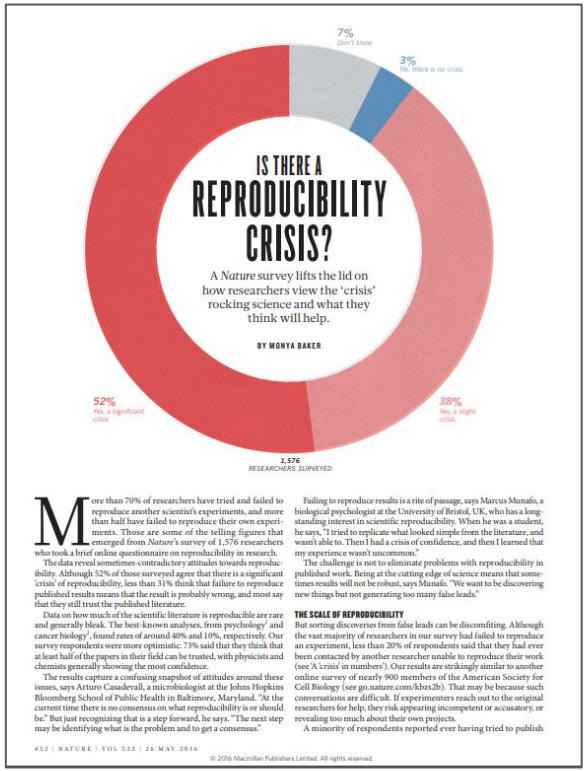
Открытые данные: FAIR-концепции

Статьи о данных (data paper)

Надежны ли журналы с открытым доступом?

Есть ли кризис воспроизводимости научных исследований?

Результаты опроса 1576 исследователей показали, что более 70% исследователей пытались и не смогли воспроизвести эксперименты другого ученого, и более половины не смогли воспроизвести свои собственные эксперименты.



IS THERE A REPRODUCIBILITY CRISIS?

A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

BY MONYA BAKER

More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from *Nature*'s survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.

The data reveal sometimes-contradictory attitudes towards reproducibility. Although 52% of those surveyed agree that there is a significant 'crisis' of reproducibility, less than 31% think that failure to reproduce published results means that the result is probably wrong, and most say that they still trust the published literature.

Data on how much of the scientific literature is reproducible are rare and generally bleak. The best-known analyses, from psychology[1] and cancer biology[2], found rates of around 40% and 10%, respectively. Our survey respondents were more optimistic: 73% said that they think that at least half of the papers in their field can be trusted, with physicists and chemists generally showing the most confidence.

The results capture a confusing snapshot of attitudes around these issues, says Arturo Casadevall, a microbiologist at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. "At the current time there is no consensus on what reproducibility is or should be." But just recognizing that is a step forward, he says. "The next step may be identifying what is the problem and to get a consensus."

Failing to reproduce results is a rite of passage, says Marcus Munafo, a biological psychologist at the University of Bristol, UK, who has a long-standing interest in scientific reproducibility. When he was a student, he says, "I tried to replicate what looked simple from the literature, and wasn't able to. Then I had a crisis of confidence, and then I learned that my experience wasn't uncommon."

The challenge is not to eliminate problems with reproducibility in published work. Being at the cutting edge of science means that sometimes results will not be robust, says Munafo. "We want to be discovering new things but not generating too many false leads."

THE SCALE OF REPRODUCIBILITY

But sorting discoveries from false leads can be discomfiting. Although the vast majority of researchers in our survey had failed to reproduce an experiment, less than 20% of respondents said that they had ever been contacted by another researcher unable to reproduce their work (see 'A 'crisis' in numbers'). Our results are strikingly similar to another online survey of nearly 900 members of the American Society for Cell Biology (see go.nature.com/kbxs2b). That may be because such conversations are difficult. If experimenters reach out to the original researchers for help, they risk appearing incompetent or accusatory, or revealing too much about their own projects.

A minority of respondents reported ever having tried to publish

© 2016 Macmillan Publishers Limited. All rights reserved.

Baker, M. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454 (2016). https://doi.org/10.1038/533452a

Мы должны ценить воспроизводимость также, как и число опубликованных статей

https://www.turing.ac.uk/research/impact-stories/changing-culture-data-science

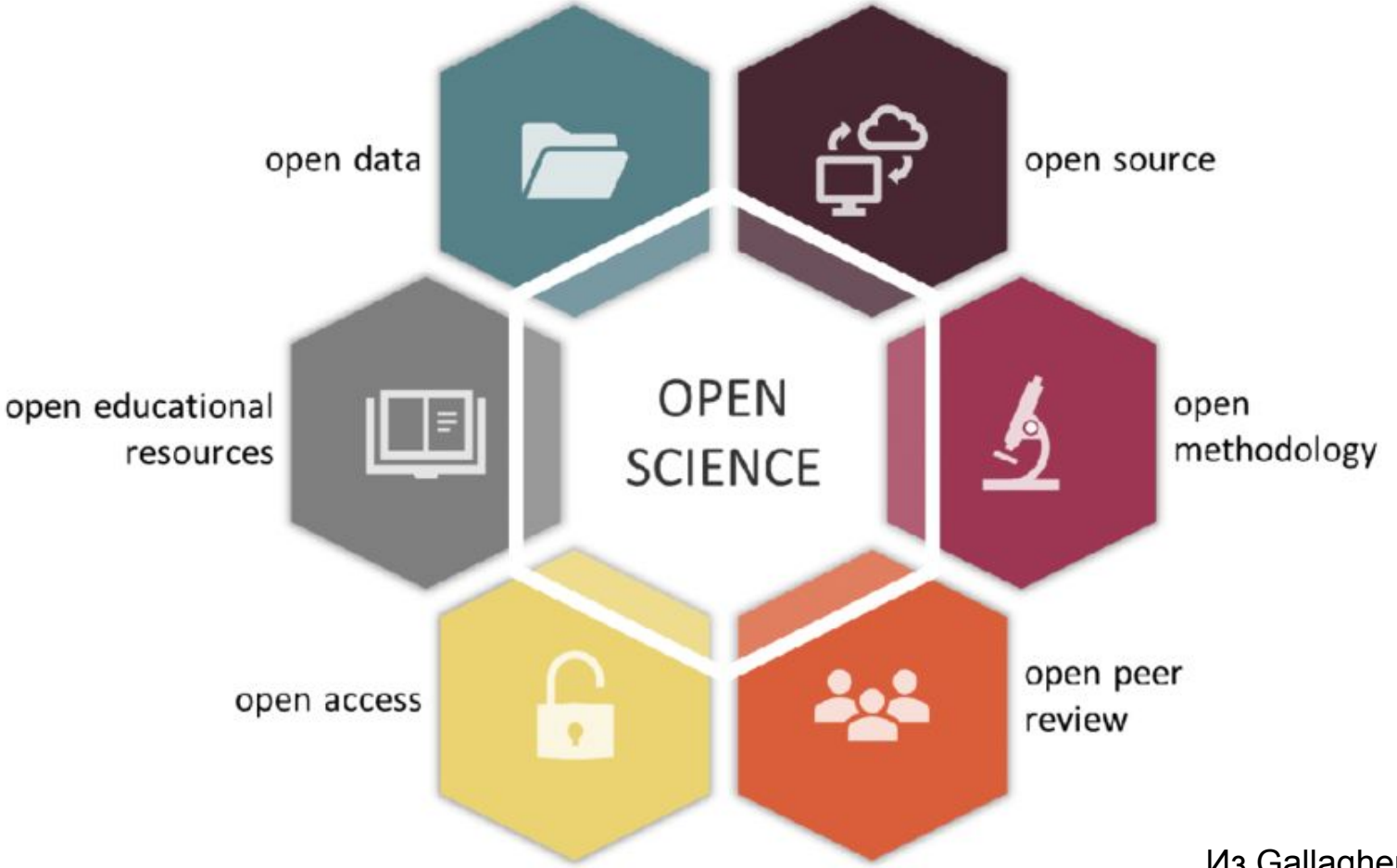
Расширение доступа к научным процессам и результатам может повысить эффективность и продуктивность научных систем за счет сокращения расходов, связанных с дублированием усилий при сборе, создании, передаче и повторном использовании данных и научных материалов, что позволит проводить больше исследований на основе одних и тех же данных.

https://www.unesco.org/en/natural-sciences/open-science

UNESCO, Проект рекомендации по открытой науке, Париж 2021

Открытый доступ: Результаты исследований, распространяемые онлайн, бесплатно и без каких-либо других препятствий - часто означают свободный доступ к исследовательским статьям.

Открытая наука: Исследователи делятся своими методами, программным кодом и данными исследований через централизованные репозитории.



Из Gallagher et al., 2019

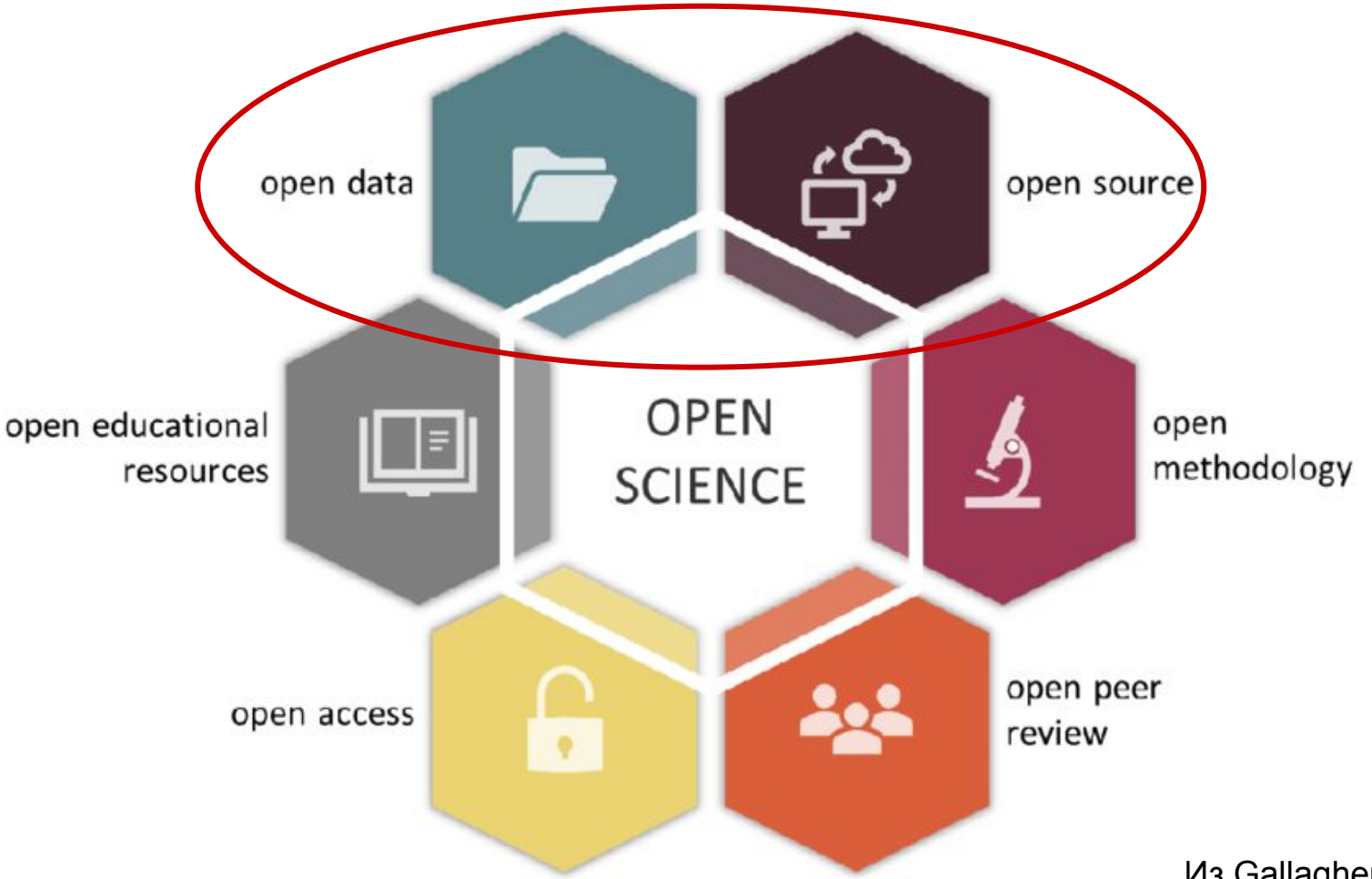
«Открытая наука» означает рамочную концепцию, которая объединяет различные движения и формы деятельности, направленные на то, чтобы сделать научные знания на различных языках открытыми, общедоступными и пригодными для всеобщего многократного использования, расширить научное сотрудничество и обмен информацией на благо науки и общества и открыть процессы создания, оценки и распространения научных знаний для социальных субъектов, не входящих в традиционное научное сообщество.

UNESCO, Проект рекомендации по открытой науке, Париж 2021

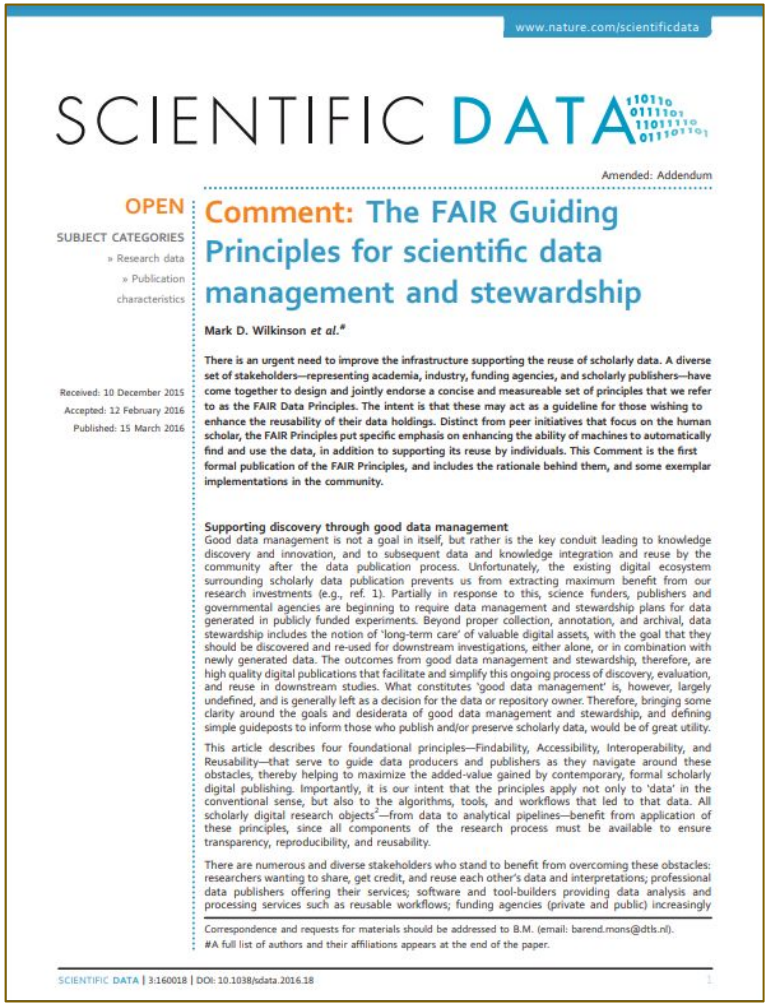
Открытые научные знания подразумевают доступ к научным публикациям, исследовательским данным, метаданным, открытым образовательным ресурсам, программному обеспечению, исходным кодам и аппаратному обеспечению, находящимся в открытом доступе или защищенным авторским правом и опубликованным на основании открытой лицензии, допускающей доступ, повторное использование, изменение целевого назначения, адаптацию и распространение на определенных условиях. Такой доступ оперативно и по возможности на бесплатной основе предоставляется всем желающим, независимо от их местонахождения, национальности, расы, возраста, пола, уровня дохода, социально-экономического положения, этапа профессиональной карьеры, дисциплины, языка, религии, инвалидности, этнической принадлежности и миграционного статуса или каких-либо других причин. Они также подразумевают возможность открытия доступа к методологиям научных исследований и процессам оценки.

UNESCO, Проект рекомендации по открытой науке, Париж 2021

https://www.turing.ac.uk/research/impact-stories/changing-culture-data-science



Из Gallagher et al., 2019



www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

Универсальная концепция, которая применяется в разных областях науки

Применение FAIR-концепции в исследованиях одобрено на саммите G20 в Ханчжоу в 2016 году

GO FAIR International Support and Coordination Office (GFISCO) - международный офис поддержки этой инициативы

https://www.nature.com/articles/sdata201618

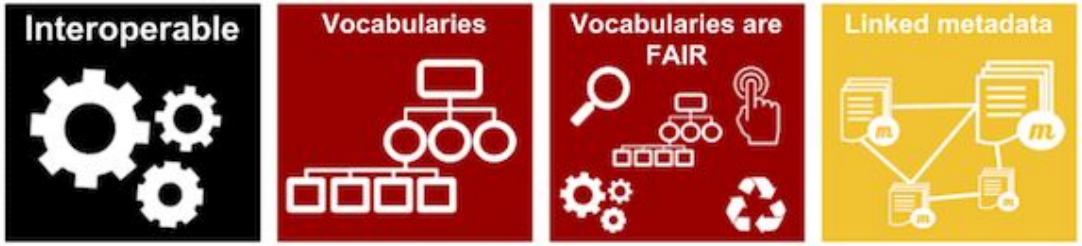

Findable
Persistent Identifiers (PIDs)
Rich metadata
Indexed data repositories
PIDs in metadata
Accessible
Standard communications protocol
Open, free protocol
Authentication, where necessary
Metadata is always available
Interoperable
Vocabularies
Vocabularies are FAIR
Linked metadata
Reusable
Metadata have multiple attributes
Usage license
Provenance
Community standards

ОБНАРУЖИМЫЕ. Данные должны иметь достаточно полные **метаданные** и уникальный **постоянный идентификатор.**

1. Data is described with rich, semantic metadata. Findable by humans and computers.
2. Data and metadata are assigned globally unique and persistent identifiers (PIDs).
3. Data and metadata is indexed in a searchable Data Catalogue, ideally DCAT-compliant.
4. Metadata includes the identifier of the data they describe.

ДОСТУПНЫЕ. Метаданные и данные **должны быть понятны** и для человека и для **компьютера**. Данные хранятся в надежном **репозитории**.

1. Data and metadata can be retrieved by their PID using standard data protocols (e.g http).
2. Data and metadata can be retrieved using open communication protocols, i.e. without proprietary tools.
3. Data access has authentication and authorisation applied where necessary.
4. Metadata should be accessible even when the data is no longer available.



СОВМЕСТИМЫЕ. Для **данных и метаданных** используется формальный, доступный и широко применимый **язык** представления знаний (стандарты и словари).

1. Data is provided in commonly understood formats, preferably open formats.
2. Metadata is defined in controlled vocabularies using knowledge representation, e.g. RDF, OWL, SKOS.
3. Metadata includes qualified references to other metadata.
4. Metadata controlled vocabularies follow FAIR principles.

МНОГОКРАТНО ИСПОЛЬЗУЕМЫЕ. Данные имеют **однозначные лицензии, описывающие правила их использования** и четкую информацию о **происхождении (данных) и любых изменениях в них**.

1. Metadata has rich, detailed attributes so a user (human or machine) can decide if the data is useful.
2. Data has a clear and accessible data usage licence (legal interoperability).
3. Data provenance is recorded, e.g. the origin of the data, any processing, any recompilation.
4. Metadata meets domain-relevant industry or community standards.

GBIF - это FAIR

Соответствует ли GBIF принципам FAIR?

ОБНАРУЖИМОСТЬ

- GBIF требует минимум **обязательных метаданных**. Все наборы данных имеют Цифровой идентификатор объекта (**DOI**).

ДОСТУПНОСТЬ

- API портала GBIF имеет **машиночитаемый интерфейс** (REST + JSON) *и использует IPT как надежное* ***хранилище данных***.

СОВМЕСТИМОСТЬ

- GBIF рекомендует применять **Язык экологических метаданных** (EML) для наборов данных и **Darwin Core** для данных о находках.

МНОГОКРАТНОЕ ИСПОЛЬЗОВАНИЕ

- GBIF требует **лицензировать данные по правилам creative commons** (CC0, CC BY, CC BY-NC). **Происхождения данных** доступно через портал GBIF, подробные сведения - через IPT.

Политика научных журналов в отношении исходных данных: обзор издательства

SPRINGER NATURE

Тип 1
Журнал призывает авторов, когда это возможно и целесообразно, размещать данные, подтверждающие результаты их исследований, в общедоступном хранилище.

Тип 2
Журнал настоятельно рекомендует, чтобы все наборы данных, на которые опираются выводы статьи, были доступны для читателей. Авторам рекомендуется предоставить информацию о доступности данных в своей статье.

Тип 3
Журнал настоятельно рекомендует, чтобы все наборы данных, на которые опираются выводы статьи, были доступны для читателей. Авторам необходимо предоставить информацию о доступности данных в своей статье.

Тип 4
Журнал требует, чтобы все наборы данных, на которые опираются выводы статьи, были доступны рецензентам и читателям. Авторам необходимо предоставить информацию о доступности данных в своей статье.

Файлы на вашем рабочем компьютере или USB-накопителе возможно содержат ценные знания, которые могут извлечь другие исследователи.

https://www.springernature.com/gp/authors/research-data

Набор данных в годовой отчет не вставишь!

DATA PAPERS

	Исследовательская статья **Research paper**	**Статья о данных** **Data paper**
ЦЕЛЬ	Проверка научной гипотезы	Описание первичных полевых данных, приведенных к требуемому формату
РАЗДЕЛЫ	Введение Материалы и методы Результаты Обсуждение Заключение	Введение Материалы и методы Описание данных
РЕЦЕНЗИРОВАНИЕ	Рецензирование текста рукописи	Аудит данных Рецензирование текста рукописи

В чем польза статей о данных*

- Регистрация приоритета и авторства данных в общепринятой научной публикации в журнале
- Публикация и цитирования статьи о данных дает авторам те же бонусы, что и публикация исследовательской статьи
- Возможность отслеживать использование и цитирование опубликованных данных
- Метаданные, опубликованные в виде статьи о данных хранятся и архивируются различными способами, что обеспечивает стабильно доступное через Интернет описание соответствующего набора первичных данных

*Chavan & Penev, 2011 https://doi.org/10.1186/1471-2105-12-S15-S2

Примеры журналов, публикующих статьи о данных*

Журнал	Издательство	Стоимость публикации	Импакт-фактор
Biodiversity Data Journal**	Pensoft	EUR 650	1.54
Scientific Data	Nature Publishing Group	EUR 1790	8.501
Taxon	IAPT	EUR 1800	2.586
Diversity	MDPI	CHF 2000	3.029
GigaScience	Oxford University Press	EUR 1089	7.658
Earth System Science Data	Copernicus GmbH	0	11.815

*полный список см. https://www.gbif.org/data-papers
**а также другие журналы этого издательства https://pensoft.net/browse-journals

PENSOFT

About Pensoft | Books | Journals | News & Blog | Contact | Natalya Ivanova

A peer-reviewed open-access journal

Biodiversity Data Journal

Making your data count! ISSN 1314-2828 (online)

Search this journal... Full Text Author Title

Submit manuscript | My drafts | My tasks

About | Articles | Topical collections | Guidelines for Authors | Editorial Team | Contacts

Most Visited Papers
Highly Accessed Papers
Most Active Editors
Most Active Reviewers
Most Productive Authors

Latest | Highly accessed


NEWS

Clarivate Web of Science Journal Impact Factor: 1,54 Q3

Scopus Scopus CiteScore 2021: 1.8 Q2

CiteScoreTracker 2022: 1.9 (updated monthly)

Become Editor | Edit a Topical Collection | Post your news

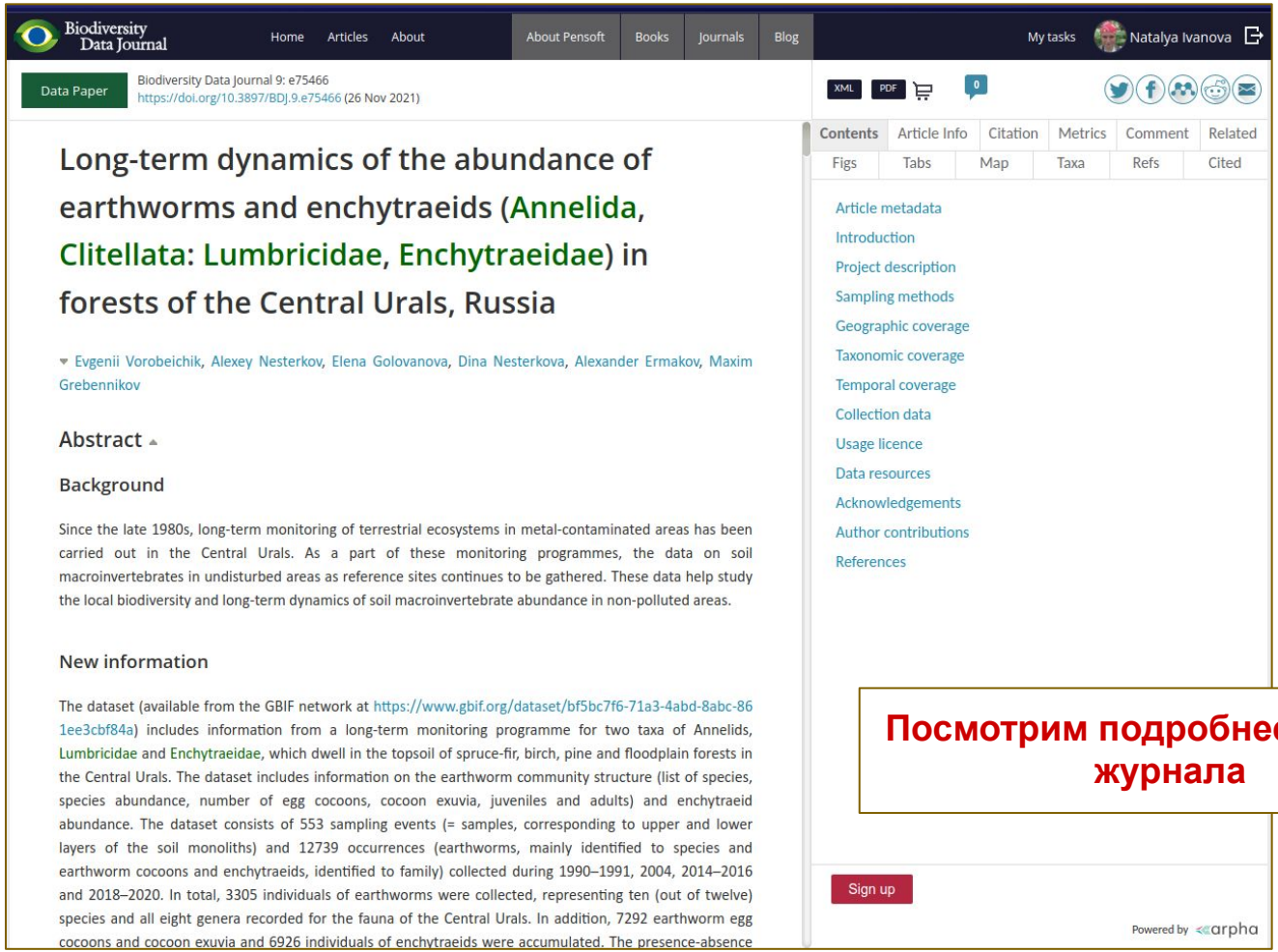
Publication types: Single Taxon Treatments, Species Conservation Profiles, Alien Species Profiles, Taxonomic Papers, Species Inventories, Identification Keys, Data Papers, Software Descriptions

Archived: PubMedCentral, CLOCKSS, Zenodo, Portico, Europe PMC

Indexed: Web of Science (SCIE), Web of Science (Current Contents), Web of Science (BIOSIS Prev.), Web of Science (Biol. Abstr.), Web of Science (Zool. Rec.), Scopus, DOAJ, AGRICOLA, AGRIS, Altmetric, ASOS Indeks, Baidu Scholar, BASE, British Library, Cabell's Directory, CABI (CAB Abstracts), ChronosHub, CNKI, CrossRef, Dimensions, EBSCO Essentials, EBSCOhost, eLibrary, EZB, GALE Academic OneFile, Genamics JournalSeek, GoOA, Google Scholar, iDiscover (University of Cambridge), IPNI, JournalTOCs, LetPub, Library of Congress, LIVIVO, MIAR, MyScienceWork, NAVER, NAVIGA, Norwegian Register, OpenAIRE, OpenCitations, Ou-Publier?, Paperity, ProQuest, ProQuest Central, ProQuest (Natural Science Collection), PubAg, Publons, PubMed/MEDLINE, QOAM (Quality Open Access Market), ReadCube, Research Gate, ROAD, Science Data Bank, ScienceOpen, Scilit, Semantic Scholar, Sherpa/Romeo, SJR Scimago, SOLO (Search Oxford Libraries Online), Transpose, Ulrichsweb™, Unpaywall, Vifabio, WanFang Data, WorldCat, ZDB

ISSN 1314-2828 (online) | ISSN 1314-2836 (print)

https://bdj.pensoft.net/



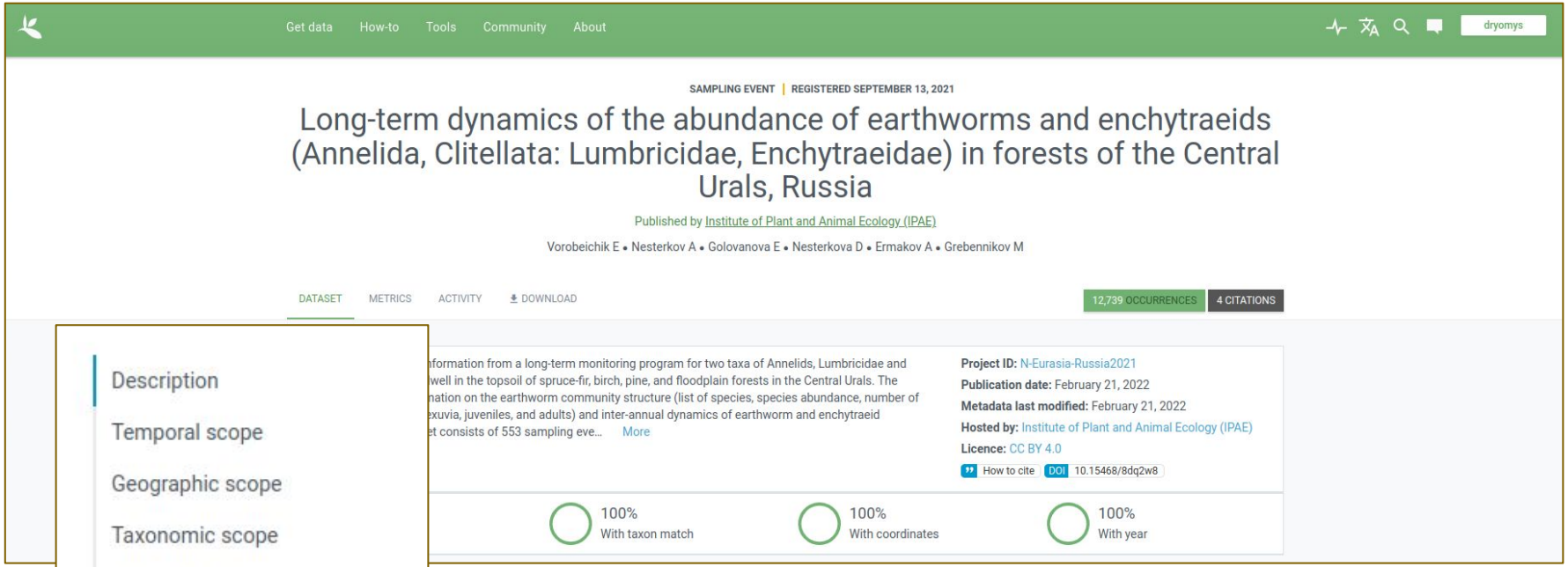
Посмотрим подробнее на сайте журнала

https://doi.org/10.3897/BDJ.9.e75466

Можно ли в одну статью включить несколько наборов данных?

Можно!

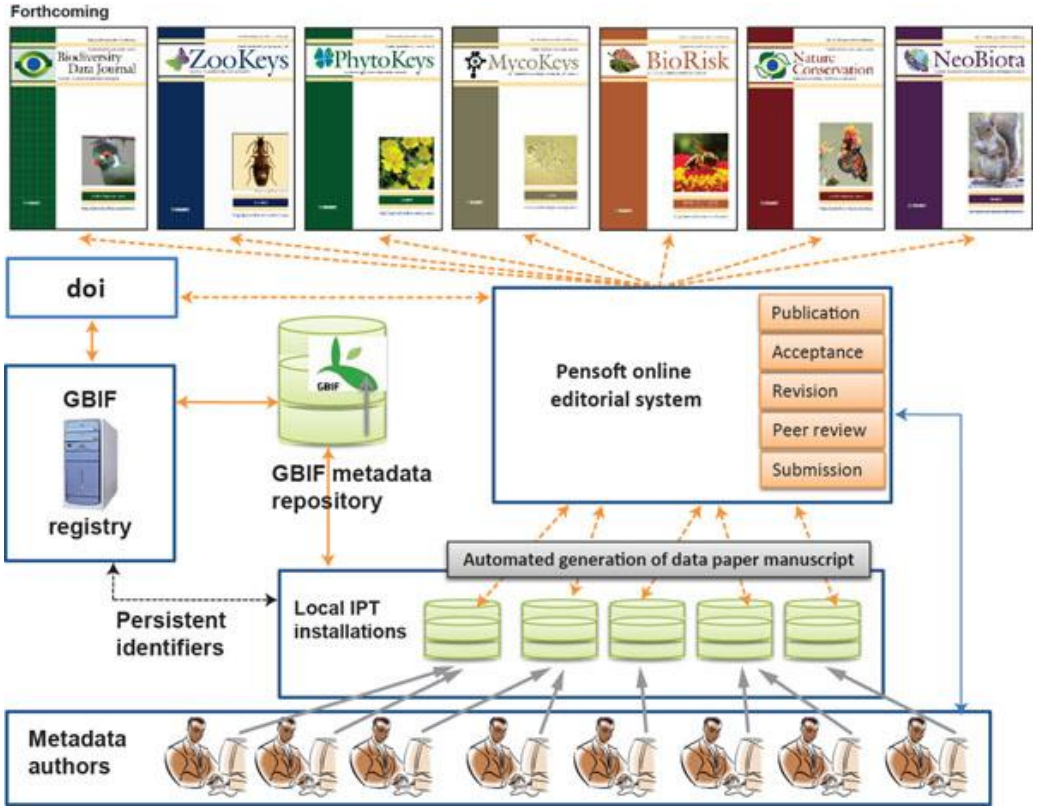
Пример
Kuznetsova N, Ivanova N (2020) Diversity of Collembola under various types of anthropogenic load on ecosystems of European part of Russia. Biodiversity Data Journal 8: e58951.
https://doi.org/10.3897/BDJ.8.e58951



Качественное описание набора данных в GBIF - 50% статьи о данных

https://www.gbif.org/dataset/bf5bc7f6-71a3-4abd-8abc-861ee3cbf84a

Взаимодействие Pensoft с порталом GBIF в процессе подготовки и публикации статьи о данных



The data paper: a mechanism to incentivize data publishing in biodiversity science

Из Chavan & Penev, 2011

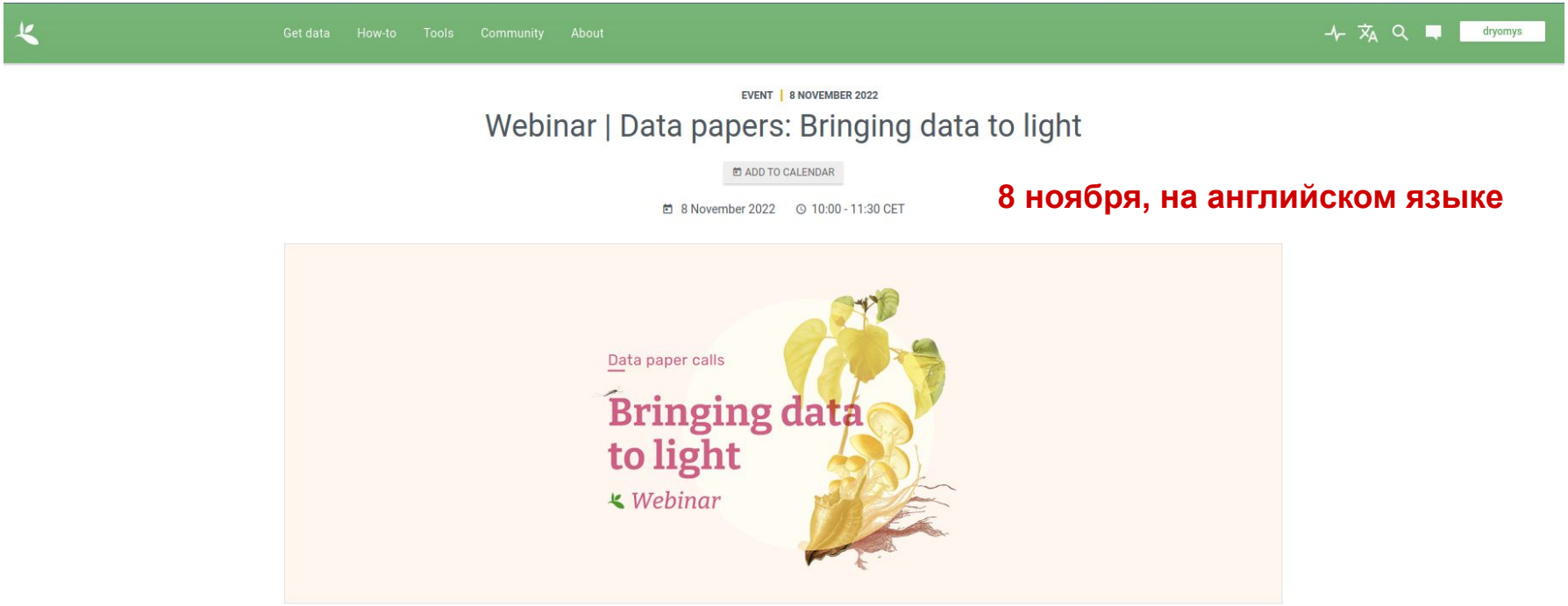
Dr. Robert Mesibov
in millipede collecting hat.
Источник

Этапы обработки статьи о данных в BDJ

- Проверка оформления рукописи на соответствие правилам журнала. Оценка качества английского языка.
- **Аудит данных**
- Рецензирование текста рукописи

- Подача рукописей **до 1 декабря**
- Набор данных должен быть опубликован в 2022 году и содержать **> 7 000 записей** о встречах видов
- Территория: Россия, Украина, Беларусь, **Казахстан**, Кыргызстан, Узбекистан, Таджикистан, Туркменистан, Молдова, Грузия, Армения, Азербайджан

https://www.gbif.org/news/B00TTCi85YIy3qNo5KXSM/call-for-data-papers-describing-datasets-from-northern-eurasia-extended

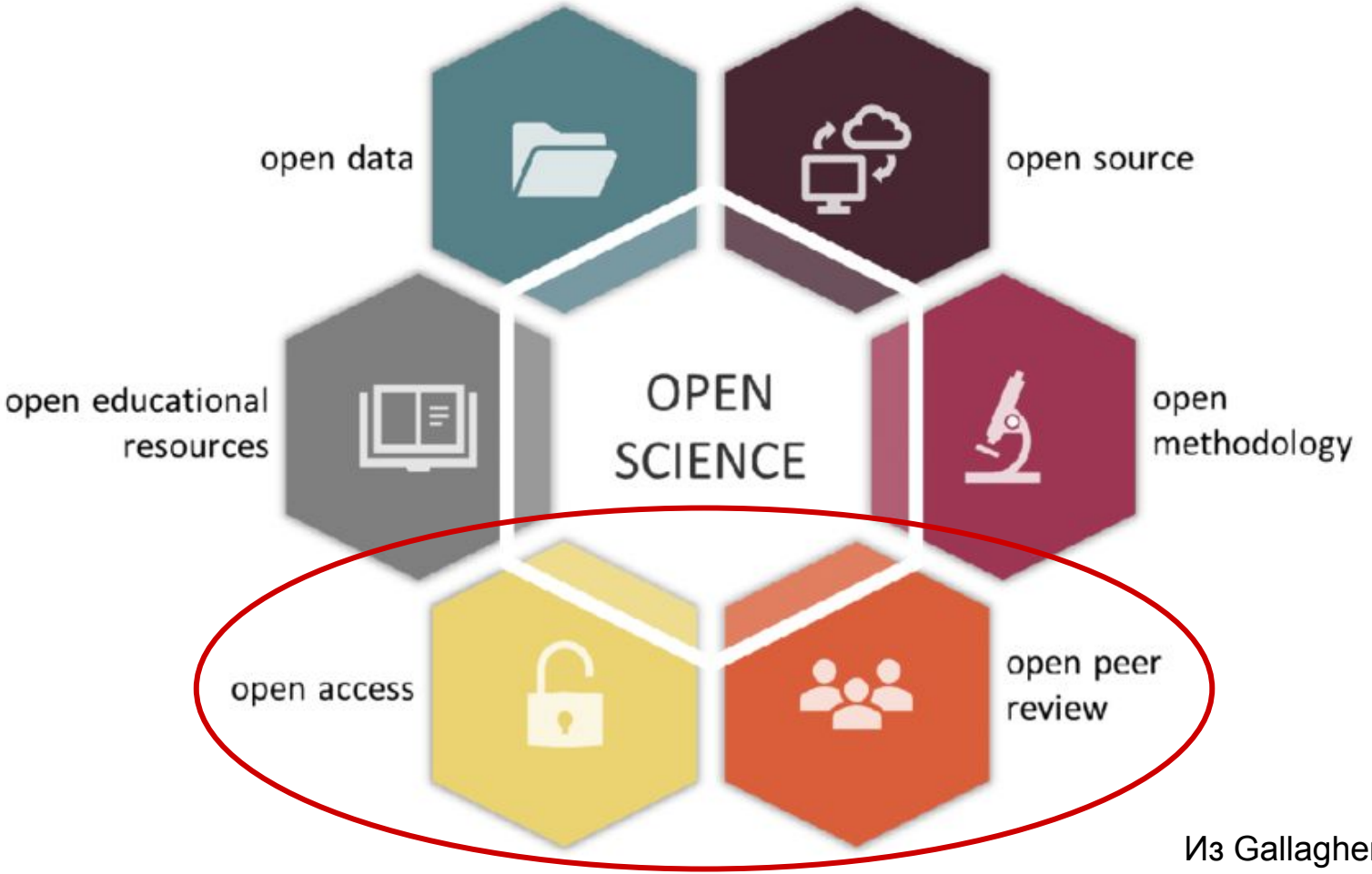


8 ноября, на английском языке

Нужно зарегистрироватся

Northern Eurasia, до 1 декабря 2022
Freshwater biodiversity, до 15 декабря 2022
Disease vectors, до 30 апреля 2023

https://www.gbif.org/event/1OaexeiM2ktptLjYoNGLUE/webinar-or-data-papers-bringing-data-to-light



Из Gallagher et al., 2019

Open Access Publishing

RESEARCHER
decides where to publish

SEE SHERPA RoMEO for
OA & self archiving options
sherpa.ac.uk/romeo

PUBLISH in an **Open access journal**

PAY Article Processing Charge (**APC**)
if required

GOLD OA ROUTE

IMMEDIATE
Open Access via publisher

PUBLISH in a **subscription-based journal**

SEARCH for
OA repository
opendoar.org

GREEN OA ROUTE

SELF ARCHIVE in a repository
based on publisher policy

IMMEDIATE or DELAYED
Open Access
(depending on publisher's policy)

HYBRID JOURNAL
subscription-based journal
With a paid open access option

PAY Article Processing Charge (**APC**)
if required

IMMEDIATE
Open Access via publisher

Darren Chase | darren.chase@stonybrook.edu Stony Brook University | Libraries

на 2018 год

DOAJ - Directory of Open Access Journals

Надежны ли журналы открытого доступа?

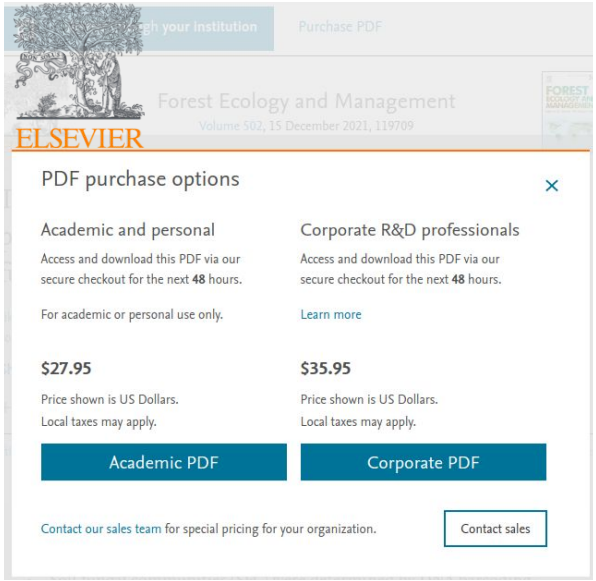
Критерии

- Редакционная коллегия. Известны ли члены редколлегии в Вашей предметной области?
- Прозрачная политика в области возникновения конфликтов интересов
- Рецензирование рукописей
- Есть ли у журнала международная читательская аудитория?
- Цитируются ли статьи, опубликованные в этом журнале?

https://authorservices.taylorandfrancis.com/are-open-access-journals-good-quality/#

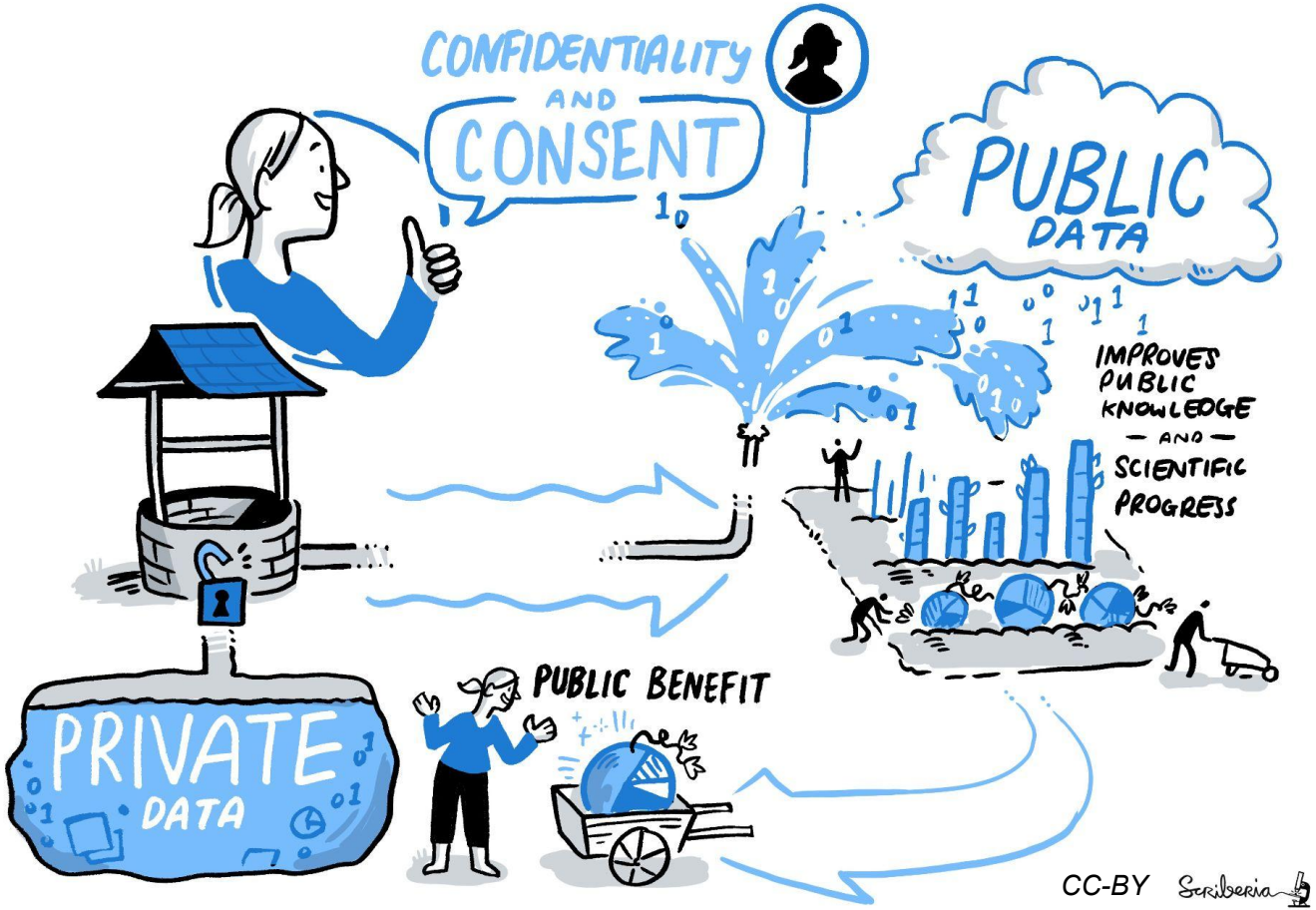
Кто платит за статью: писатель или читатель?

Журналы с открытым доступом плохие потому что публикуют статьи за деньги



Open Access

Публикация в рейтинговом журнале - основание для нового гранта



DOI: 10.5281/zenodo.3695300

Материалы для самостоятельного изучения

De Prins J. (2019) Global Open Biodiversity Data: Future Vision of FAIR Biodiversity Data Access, Management, Use and Stewardship. Biodiversity Information Science and Standards, 3: e37190 DOI: 10.3897/biss.3.37190

Wilkinson M.D., Dumontier M., Jan Aalbersberg IJ., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3: 160018. DOI: 10.1038/sdata.2016.18

Chavan, V., Penev, L. (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics 12 (Suppl 15), S2. https://doi.org/10.1186/1471-2105-12-S15-S2