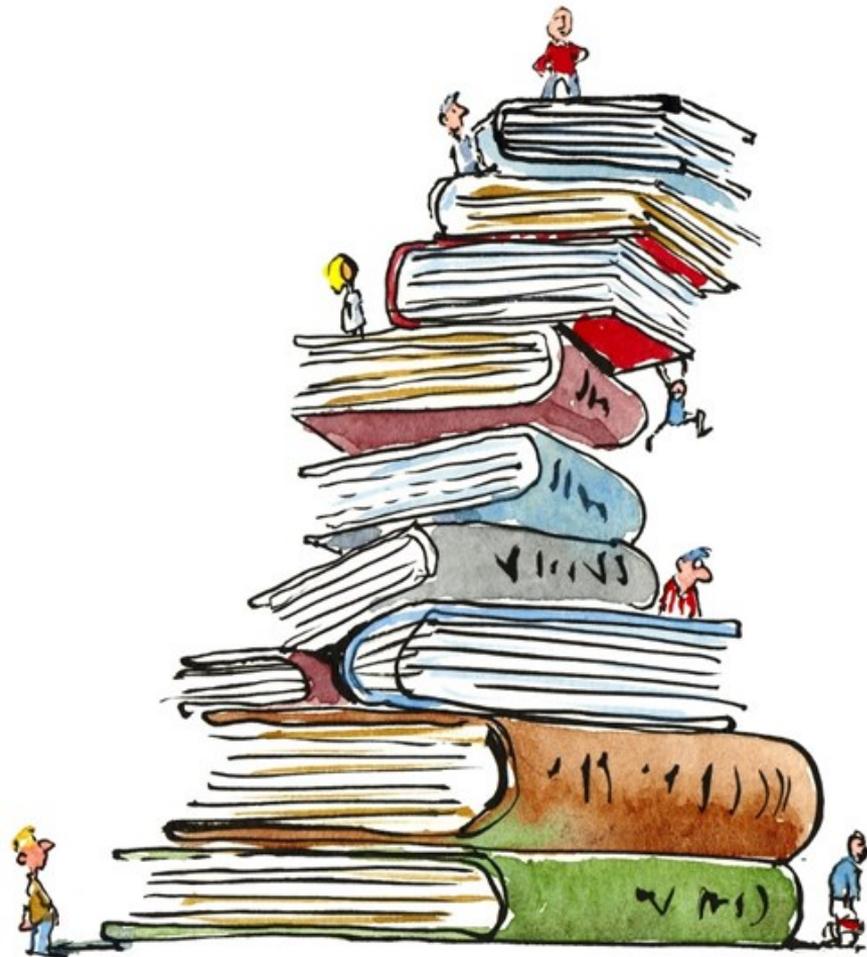


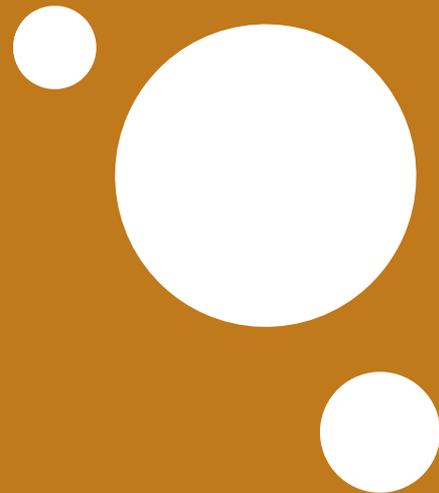
# Наукометрия на открытых данных

апрель 2023





# 1.Эволюция данных



# Указатели цитирования

Статья  $p_0$  цитирует статьи  $[p_1-p_n]$

## Ограничения

- зависимость от качества работы издательства
  - 30% в топ-10 LIS журналах ([1992](#))
  - 4-40% в топ-5 мед. журналах ([2000](#))
  - 38% в биомед. журналах ([2007](#))
  - 40% в журнале RQES ([2021](#))
  - <2% в 100 Nursing журналах ([2022](#))
- тысячи стилей оформления библиографии (<https://citationstyles.org> 10К+ стилей)
- доступность данных ([I4OC](#)) - с 03.06.2022 открытая лицензия для всех, но свыше 10 тыс. издателей не предоставляют списки литературы

Specialty	Sample year	# of journals studied	Sample size (# of references) per journal	Error rates (% of references containing at least 1 error)
Anesthesiology	1990, 1994	1	3,967, 2,183	32.0%, 41.0%
Ophthalmology	2003	10	200	16.0% combined
Pediatric surgery	2001	3	903, 318, 285	31.2%, 37.9%, 34.6%
General medicine	1984	6	50 each	24.0% combined
Public health	1986	3	50 each	31.0% combined
Surgery	1987	3	50 each	48.0% combined
Otolaryngology	1997	4	41, 42, 44, 41	34.1%, 28.6%, 31.8%, 56.1%
Emergency medicine	1991	3	145 combined	27.5%
Radiology	1993	2	47, 48	27.0%, 45.0%
Critical care	1993	1	96	35.6%
Critical care nursing	2000	3	48, 118, 78	29.0%, 20.3%, 23.0%
Obstetrics and gynecology	1995	3	50 each	57.8%, 60.9%, 66.7%
General medicine	1999	5	395, 280, 213, 317, 352	4.1%, 6.1%, 17.4%, 27.7%, 40.3%
Allergy and immunology	1999	3	788, 589, 410	22.1%, 30.4%, 28.0%

# Идентификаторы объектов (PID)

Статья  $p_0$  цитирует статьи  $[p_1-p_n]$  и у всех  $p$  из  $\{P\}$  есть уникальный идентификатор  $ID_p$   
Аналогично, создаются  $ID_x$  для изданий ( $x=S$ ), авторов ( $x=A$ ), организаций ( $x=O$ ), ...

Who has committed to the POSI principles?

These organizations or initiatives (listed alphabetically) have formally adopted the POSI principles by publishing an initial self-audit, and committed to routinely demonstrating evidence of following POSI in practice.

- Crossref: [POSI fan tutte](#) (2022-March-08) and [original](#) (2020-December-02)
- [CORE](#) (original posted 2022-May-23)
- [DataCite](#) (original posted 2021-August-30)
- [DOAJ](#) (original posted 2022-October-06)
- [Dryad](#) (original posted 2020-December-08)
- [Europe PMC](#) (original posted 2022-February-21)

## Ограничения

- Журналы используют PID по своему усмотрению.
- Часть популярных PIDs генерируются закрытыми системами (ResearcherID, ISNI, VIAF, Scopus Author ID, Ringgold, SPIN-код, ... )
- Некоторые PIDs являются лишь условно “открытыми” (ISSN, DOI, ORCID, ROR, Wikidata и т.д.)
- Многие декларируют приверженность “принципам открытой академической инфраструктуры” (POSI), но на практике вынуждены соблюдать другие правила.

# Онтологии

Иерархии и атрибуты для объектов и их отношений (например, вклад авторов, типы ключевых слов, положение цитирования в разделе статьи, и т.д.).

**Eczema** / Chemically Induced

**Metals, Heavy** / toxicity

Age: Newborn

Sex: Female

Высокая специфичность поисковых запросов!

Самоцитирование на уровне авторов, журналов, организаций.

## Ограничения

- Отсутствие универсальной онтологии (зато есть множество предметных)
- Гармонизация изменений между базами данных
- Высокая стоимость подготовки (разметки) данных
- Низкая квалификация пользователей БД ([из 2.9М запросов к PubMed/MEDLINE, лишь 3% содержали два или более тэга, 75% пользователей не использовали их](#))

# Графы знаний

Объединение нескольких (в идеале, многих) онтологий через общие идентификаторы

- Женщины-ученые, которые имеют научные награды и в своих публикациях цитировали статьи из журнала “Молекулярная биология”? <https://w.wiki/5r9Y>
- Страны без выхода к морю, но граничащие со странами, имеющими выход <https://w.wiki/6cxJ>
- Все библиотеки Австрии <https://w.wiki/6cxM>
- Университеты Москвы на расстоянии не более 500 метров от станций метрополитена <https://w.wiki/6d39>

## Ограничения

- Неполнота (в Wikidata каждую неделю предлагаются новые свойства!)
- Обновление
- Требования к квалификации исследователей (напр. SPARQL)
- Требования к инфраструктуре (напр. федеративный поиск)

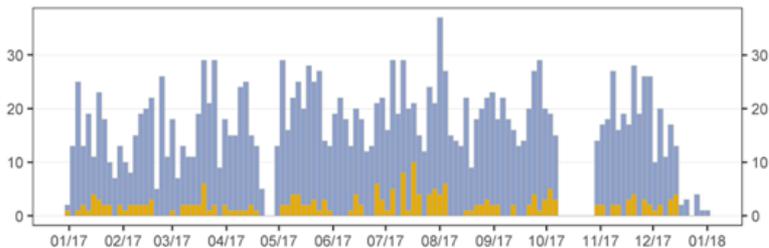
# Wikidata и SPARQL запросы

Для исследования Викиданных используют запросы на языке SPARQL, которые отправляют через веб-интерфейс [Wikidata Query Service](#) или с помощью API (R: [WikidataR](#), Python: [Wikidata](#), [другие примеры](#)).

- Перечень сервисов, имеющих SPARQL и ID в Wikidata <https://w.wiki/6сха> (крупнейшие - GND, BnF, UniProt, Open Citations и др.)
- Примеры простых SPARQL-запросов  
[https://www.wikidata.org/wiki/Wikidata:SPARQL\\_query\\_service/queries/examples](https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples)
- Примеры более сложных SPARQL-запросов  
[https://www.wikidata.org/wiki/Wikidata:SPARQL\\_query\\_service/queries/examples/advanced](https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples/advanced)
- Коллекция запросов <https://www.wikidata.org/wiki/User:MartinPoulter/queries>
- Архив примеров SPARQL-запросов к Wikidata (!!!)  
[https://www.wikidata.org/wiki/Wikidata:SPARQL\\_query\\_service/qotw](https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/qotw)
- О профилях организаций: <https://openriro.github.io/posts/wikidata-profile/>
- О профилях журналов <https://podpiska.rfbr.ru/materials/wikidata4journals>
- Курс “Программирование Викиданных” <https://w.wiki/62E>

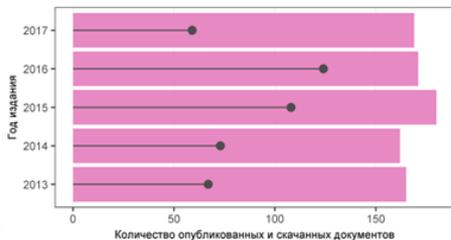
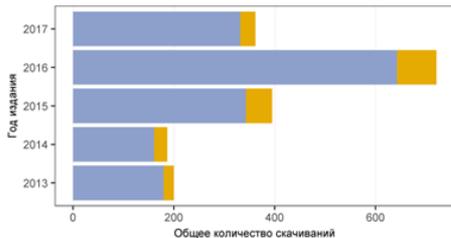
# Смешанные данные (Sci-Hub 2017)

Интенсивность скачиваний (1 столбец = 3 дня)

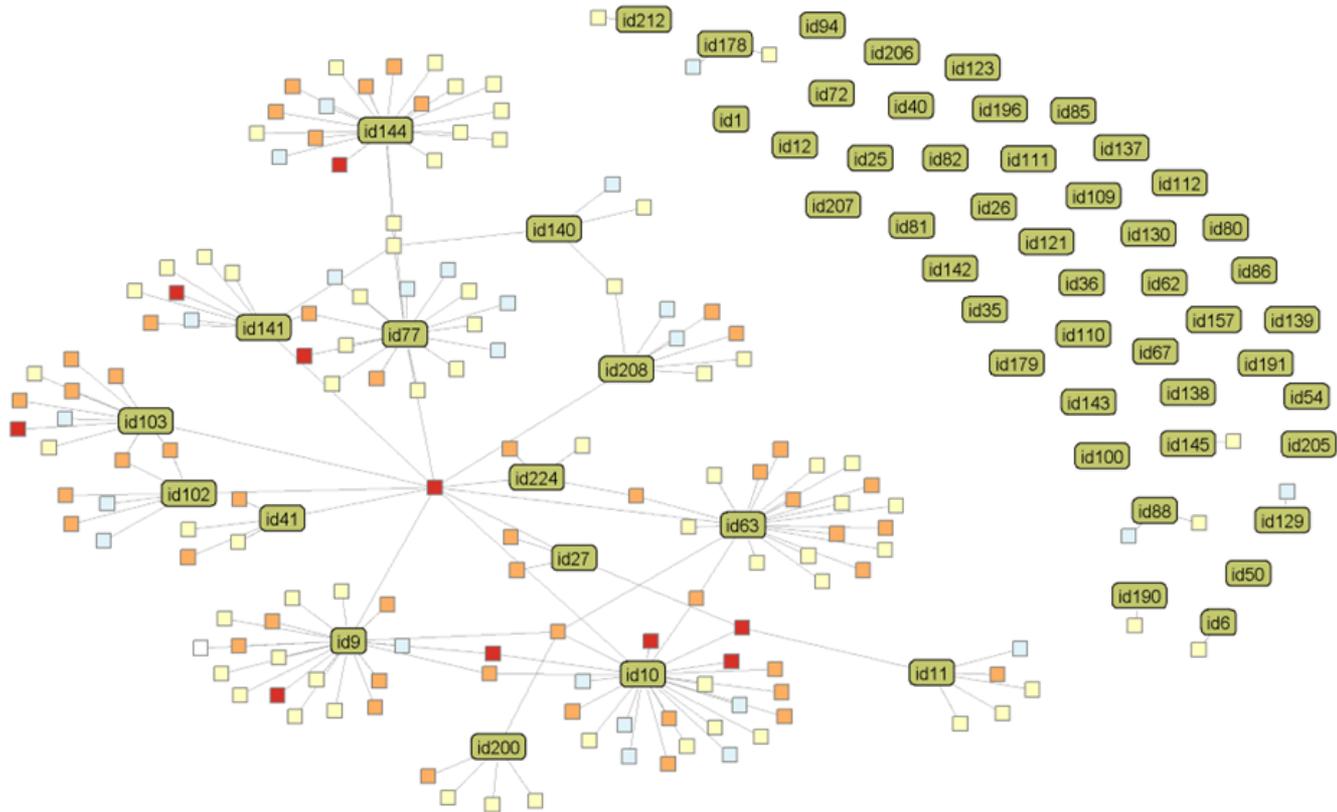


**Источник**  
■ из-за рубежа  
■ из России

	Документов	Скачиваний		Документов	Скачиваний
China	256	642	Germany	29	34
Russia	144	192	South Korea	26	28
United States	78	89	Hong Kong	22	29
India	68	91	Portugal	22	26
Brazil	54	86	Turkey	19	23
France	34	43	Poland	17	21
Spain	32	47	Argentina	17	19
Taiwan	32	38	Chile	15	23
Mexico	30	34	Italy	15	17
Iran	29	42	Canada	14	19



# Смешанные данные (Twitter)



# Смешанные данные (OpenRIRO)

Table 9 - Scival

Table 10 - Russian Universities Assessment System (by the Ministry of Science & Higher Education)

Table 11 - Web of Science

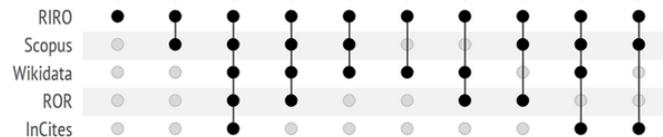
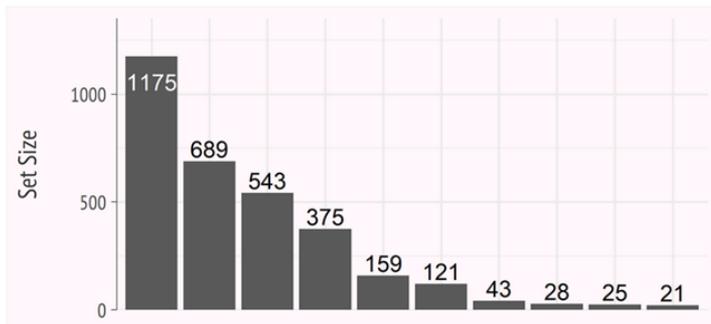
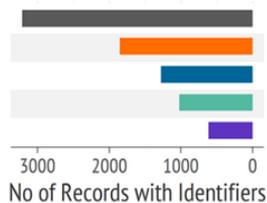
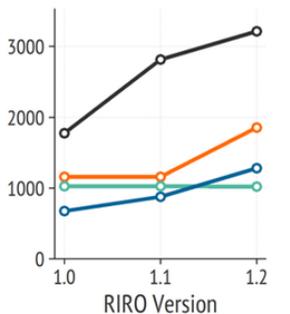
Table 12. eLIRBARY.ru

All IDs for 3 organizations

Feedback

[Show code](#)

## Organizations in RIRO and Other Identifiers (head organizations only)



Date: 09/15/2021  
Source: RIRO v.1.2

# Смешанные данные (Белый список)



РОССИЙСКИЙ ЦЕНТР  
НАУЧНОЙ  
ИНФОРМАЦИИ

## ACM TRANSACTIONS ON DATABASE SYSTEMS

0362-5915, 1557-4644

Scopus WoS CC DBLP Inspec Compendex

Основная информация

Показатели

Категории

Квартили

Списки

Ссылки

Анализ

### ОСНОВНАЯ ИНФОРМАЦИЯ

ACM TRANSACTIONS ON DATABASE SYSTEMS

Заголовок на англ.

0362-5915 (Scopus, Web of Science), 1557-4644 (Web of Science)

ISSN

Английский

Язык

Crossref

Агентство регистрации DOI

-

Заголовок на рус.

США (Web of Science, ISSN.org, Scimago JR)

Страна

### ПРОФИЛИ ЖУРНАЛА

- Sherpa Romeo
- Scilit
- OpenAlex
- OpenAlex API
- Wikidata
- Scholia
- DBLP
- Fatcat



## 2. Качество метаданных



# Исследования качества

**to err is human?**

Статей о качестве метаданных в разных БД превеликое множество.

- **German Research Council (DFG) → California Department of Fish and Game**
- **CGS → Canadian Geriatrics Society**, в оригинале **China Geological Survey**
- **ISF → Iowa Science Foundation**, в оригинале **Israel Science Foundation**
- **NSF** (в 1 примере) → **National Sleep Foundation**
- **FANO → Captain Fanourakis Foundation**
- **RSCF → Richmond County Savings Foundation** (в 2019) → **Robert Sterling Clark Foundation** (2023) - статья в УФН (2018)

Но все же чаще виноваты издатели, которые попросту не включают сведения в пакет метаданных, используют дикие схемы данных в разметке, сливают строки аффилиаций или оборачивают сведения о финансировании в виньетки и т.д.

# Исследование 1

## Цитируемость статей

Базы данных: SC = Scopus, WS = Web of Science Core Collection, CR = CrossRef, LN = Lens, S2 = Semantic Scholar, OA = OpenAlex  
Массив статей: 58634 публикаций 2016-2021 гг. с участием исследователей из РФ.



© РФФИ, иллюстрация, 2022  
Дата создания: январь-февраль 2022

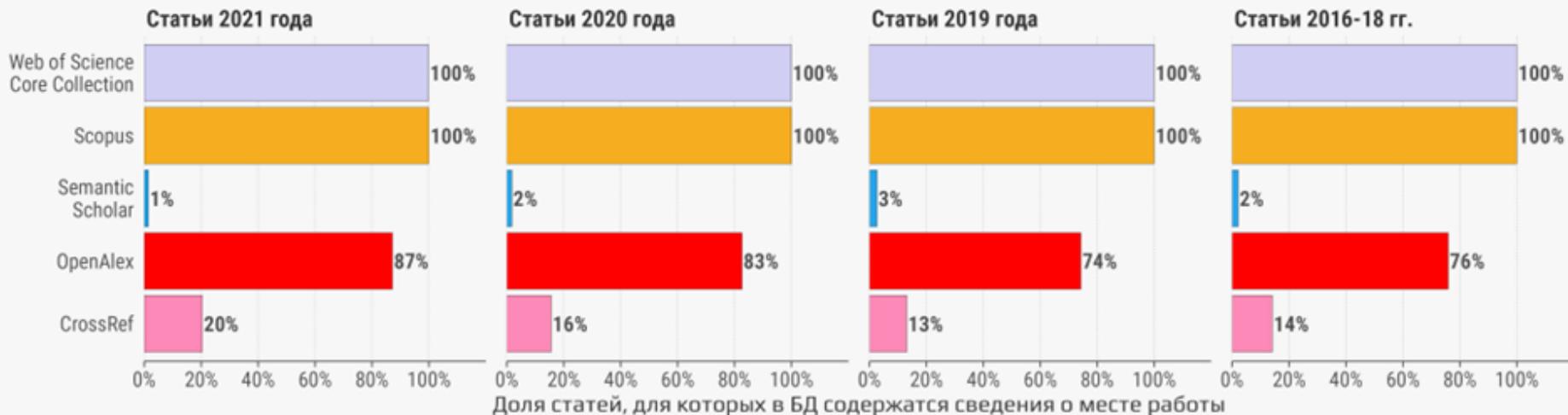
[https://podpiska.rfbr.ru/storage/reports2021/2022\\_meta\\_quality.html](https://podpiska.rfbr.ru/storage/reports2021/2022_meta_quality.html)

# Исследование 1

## Наличие информации о месте работы авторов

Базы данных: SC = Scopus, WS = Web of Science Core Collection, CR = CrossRef, S2 = Semantic Scholar, OA = OpenAlex

Массив статей: 57570 публикаций 2016-2021 гг. с участием исследователей из РФ. Исключены статьи с 100 и более авторов.



© РФФИ, иллюстрация, 2022  
Дата создания: январь-февраль 2022

[https://podpiska.rfbr.ru/storage/reports2021/2022\\_meta\\_quality.html](https://podpiska.rfbr.ru/storage/reports2021/2022_meta_quality.html)

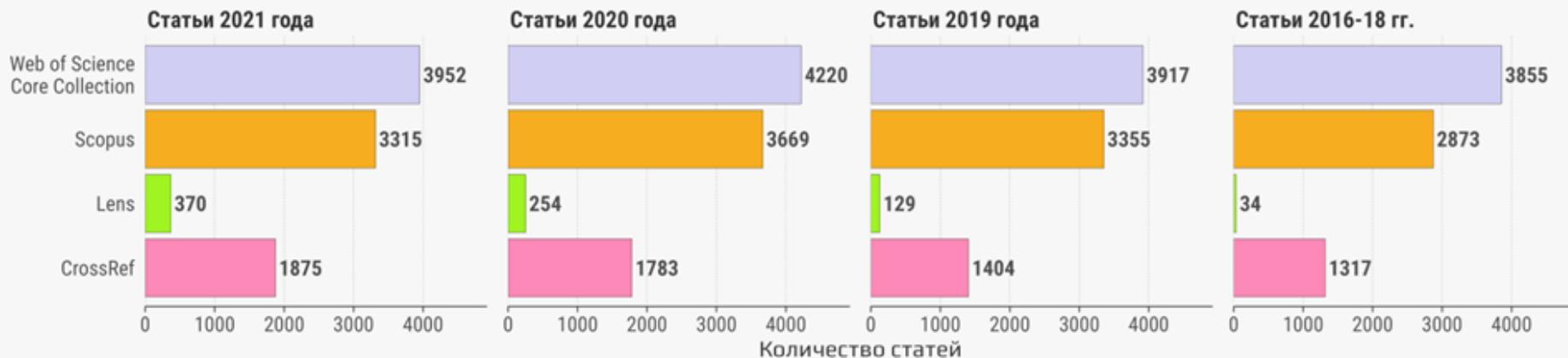
# Исследование 1

## Количество статей, содержащих сведения о гранте по маске поиска

Базы данных: SC = Scopus, WS = Web of Science Core Collection, CR = CrossRef, LN = Lens

Массив статей: 58634 публикаций 2016-2021 гг. с участием исследователей из РФ.

Маска поиска: (№1) RFBR|RSF|Russian Foundation for Basic Research|Russian Science Foundation, (№2) [0-9]{2}-[0-9]{2}-[0-9]{5}



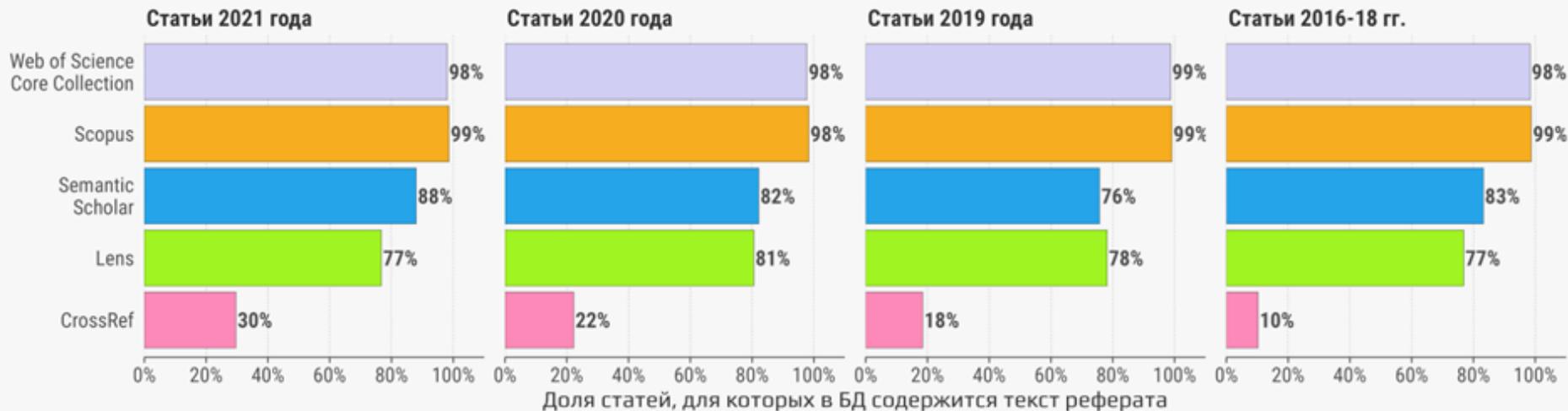
© РФФИ, иллюстрация, 2022  
Дата создания: январь-февраль 2022

[https://podpiska.rfbr.ru/storage/reports2021/2022\\_meta\\_quality.html](https://podpiska.rfbr.ru/storage/reports2021/2022_meta_quality.html)

# Исследование 1

## Наличие реферата в метаданных

Базы данных: SC = Scopus, WS = Web of Science Core Collection, CR = CrossRef, LN = Lens, S2 = Semantic Scholar  
Массив статей: 58634 публикаций 2016-2021 гг. с участием исследователей из РФ.



© РФФИ, иллюстрация, 2022  
Дата создания: январь-февраль 2022

[https://podpiska.rfbr.ru/storage/reports2021/2022\\_meta\\_quality.html](https://podpiska.rfbr.ru/storage/reports2021/2022_meta_quality.html)

# Исследование 1

## Присутствие ORCID в доступных метаданных публикаций

Базы данных: SC = Scopus, WS = Web of Science Core Collection, CR = CrossRef, LN = Lens, S2 = Semantic Scholar, OA = OpenAlex  
Массив статей: 57569 публикаций 2016-2021 гг. с участием исследователей из РФ. Исключены статьи с 500 и более авторов.



© РФФИ, иллюстрация, 2022  
Дата создания: январь-февраль 2022

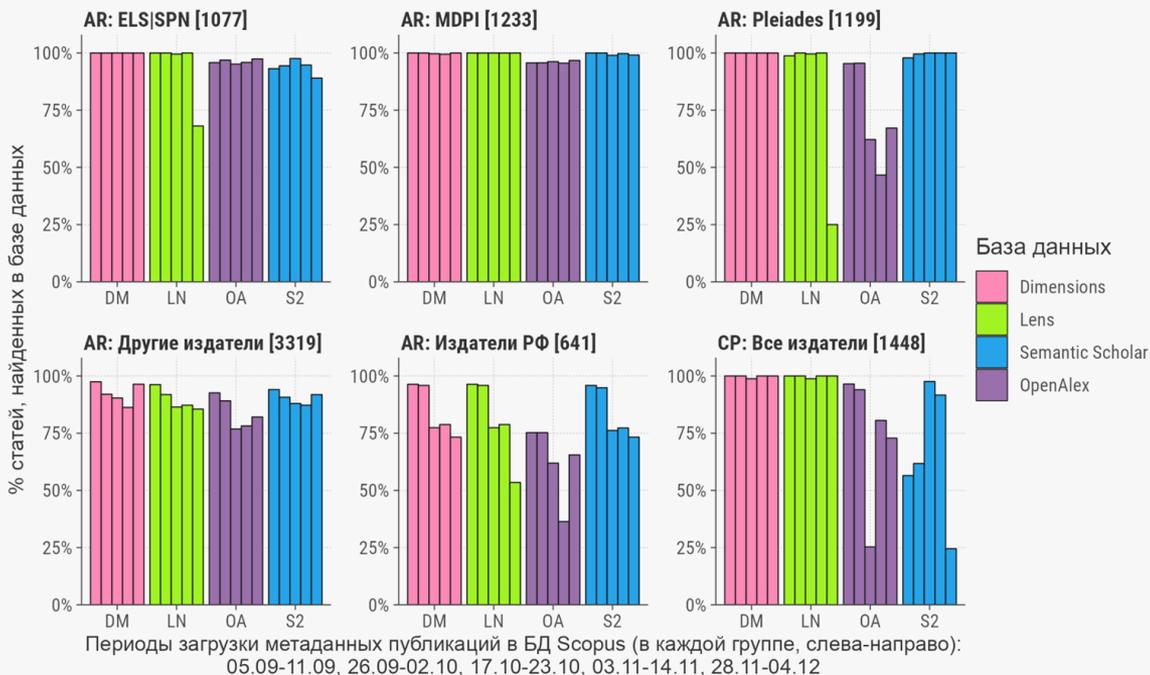
[https://podpiska.rfbr.ru/storage/reports2021/2022\\_meta\\_quality.html](https://podpiska.rfbr.ru/storage/reports2021/2022_meta_quality.html)

# Исследование 2

## Скорость индексации в Scopus и в других базах данных

Массив: 8,917 публикаций 2020-2023 гг., проиндексированных в Scopus в диапазоны дат: (05.09-11.09.2022), (26.09-02.10.2022), (17.10-23.10.2022), (03.11-14.11.2022), (28.11-04.12.2022)

AR - article/review, CP - conference paper, ELSJSPN - издания Elsevier или Springer Nature



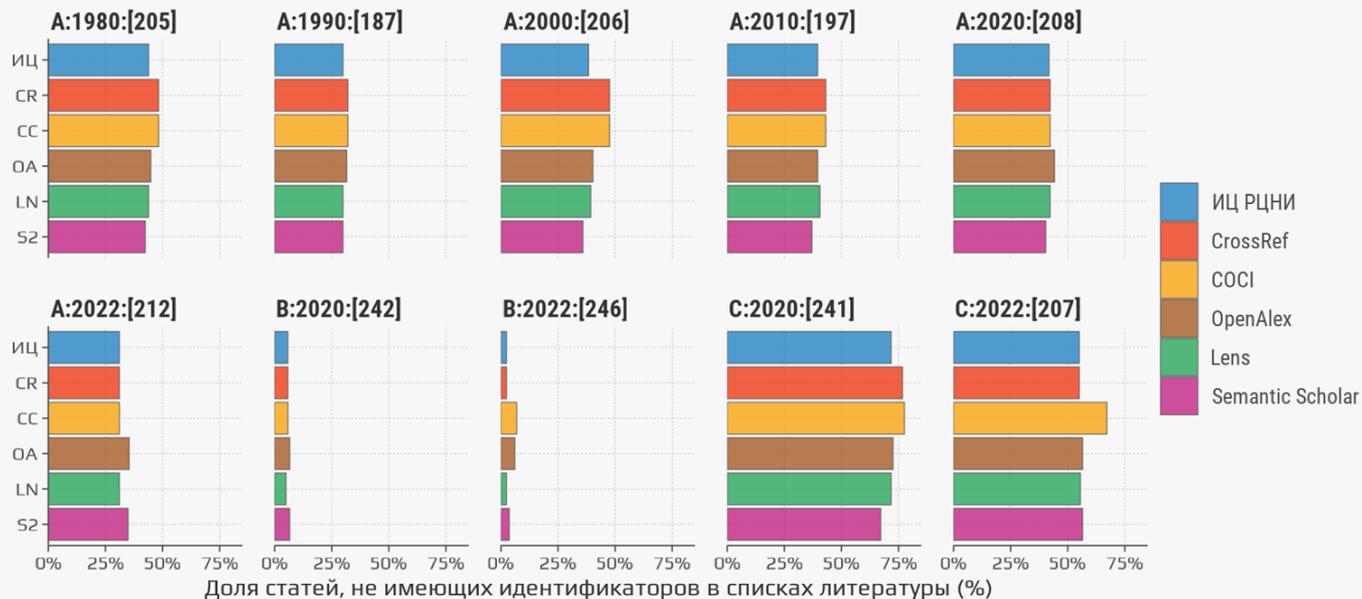
© РЦНИ, иллюстрация, 2022  
Дата создания: 18.12.2022

[https://podpiska.rfbr.ru/materials/2022\\_open\\_search\\_solutions/](https://podpiska.rfbr.ru/materials/2022_open_search_solutions/)

# Исследование 3

## Отсутствие идентификаторов в списках пристатейной литературы

Группы А-С содержат DOI из журналов разных издательств. Название группы = префикс:год публикации:[количество DOI].  
Префикс А - статьи из журналов крупных зарубежных издательств (Elsevier, Springer Nature, Wiley и др.)  
Префикс В - статьи из журналов издательств открытого доступа (PLOS, Frontiers, MDPI и др.)  
Префикс С - статьи из журналов крупных российских научных издательств.

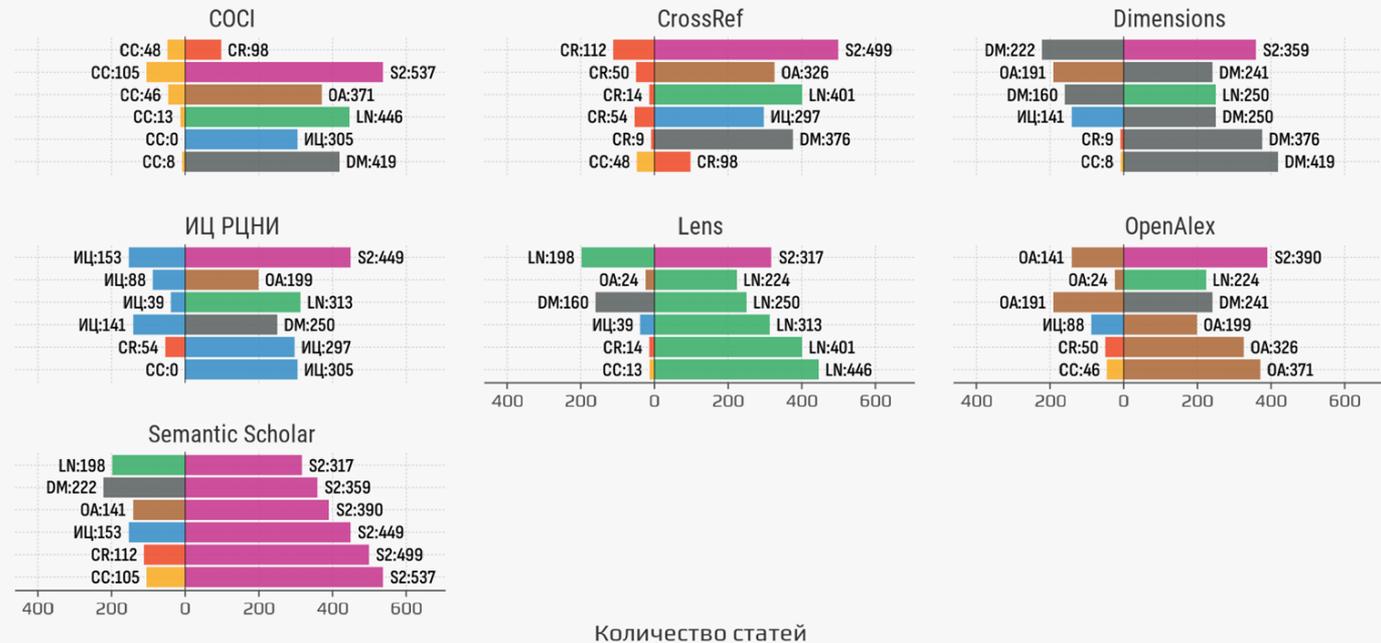


© РЦНИ, иллюстрация, 2023  
Дата создания: 31.03.2023

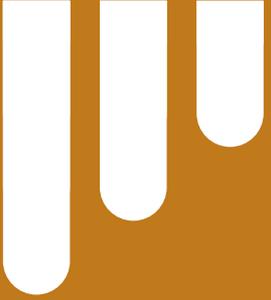
# Исследование 3

## Сравнение БД по цитируемости статей

Подпись рядом со столбцом содержит код БД (CC = COCI, CR = CrossRef, DM = Dimensions, ИЦ = ИЦ РЦНИ, LN = Lens, OA = OpenAlex, S2 = Semantic Scholar) и количество статей из массива (всего - 2151), для которых цитируемость в указанной БД выше, чем в БД сравнения.



© РЦНИ, иллюстрация, 2023  
Дата создания: 31.03.2023



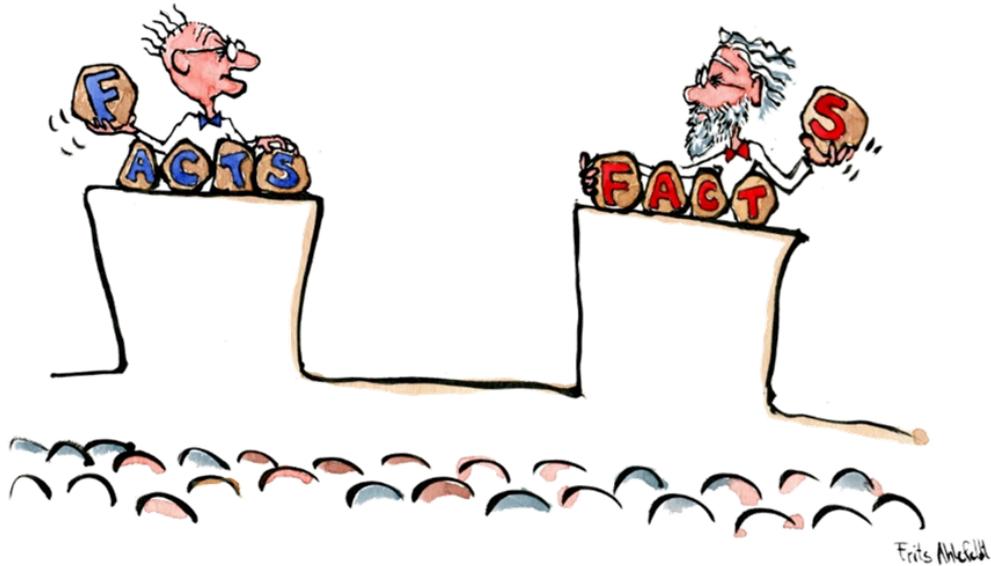
# 3. Что делать?



# Хочешь сделать хорошо? Позови хороших людей



# Будь ГОТОВ



# Помни о людях



By Frits Aeliefeldt



By HikingArtist.com



HikingArtist

# Делись с другими

- Массив данных – в репозиторий (Figshare, Zenodo, etc)

## Недостаточно круто?

- Код или пакет на GitHub
- Связанные данные в Wikidata ([QuickStatements](#) – из CSV!)

QuickStatements English  [New batch](#) [Last batches](#) [Chat](#) [Git](#) [Help](#) L

## Last batches

If you are an admin and want to stop a batch, first [Log in!](#)

#	User	Name	Status	Last change	Actions
#168356	Akbarali <a href="#">[Batches]</a>	ml label set 2 ( 25.4.23)	Running <span>DONE:7445</span> <span>INIT:2553</span> <span>RUN:1</span>	2023-04-25 20:59:29	<a href="#">Discuss/revert batch</a>

# Отдавай должное

- @science\_policy
  - @begtin
  - @HQhse «Выше квартилей»
  - @kvartil
  - @ciase\_eu
  - @scientometrics\_and\_Research\_Eval
  - @lib\_os
  - @psalchannel
  - @idscience\_ru
  - @colab.ws
  - @ANRIRUS
  - @neicon
  - @elibrary22
- а также всем энтузиастам, просветителям  
и небезразличным людям.



Frits Ahlefeldt-Laurvig  
[hikingartist.com](http://hikingartist.com)  
[fritsahlefeldt.com](http://fritsahlefeldt.com)

The illustrations protected by copyright have been purchased @ [fritsahlefeldt.com](http://fritsahlefeldt.com). Some illustrations are licensed under CC-BY-ND 3.0 (see)