Новая наукометрия 1: источники данных

Краткий обзор

28 сентября 2022 г.

НИУ ВШЭ

Иван Стерлигов

isterligov@hse.ru

План семинара

О каких данных речь?

a. **Метаданные**:
   - Общенаучные: CrossRef, OpenCitations\COCI, OpenAlex, Semantic Scholar, The Lens, Wizdom.ai, Dimensions, FATCAT, scilit.net, BASE
   - Тематические: DBLP, ADS, MEDLINE, INSPIRE HEP.
   - Метаданные источников

b. **Полные тексты**:
   - Инфраструктура: Unpaywall, DOAJ
   - Базы полных текстов: CORE, PubMed Central, препринты, Cyberleninka
   - Датасеты: Zenodo, Figshare

c. **Интеграция и майнинг**
   - Wikidata, OpenAIRE
   - The General Index, Scite, Connected Papers etc

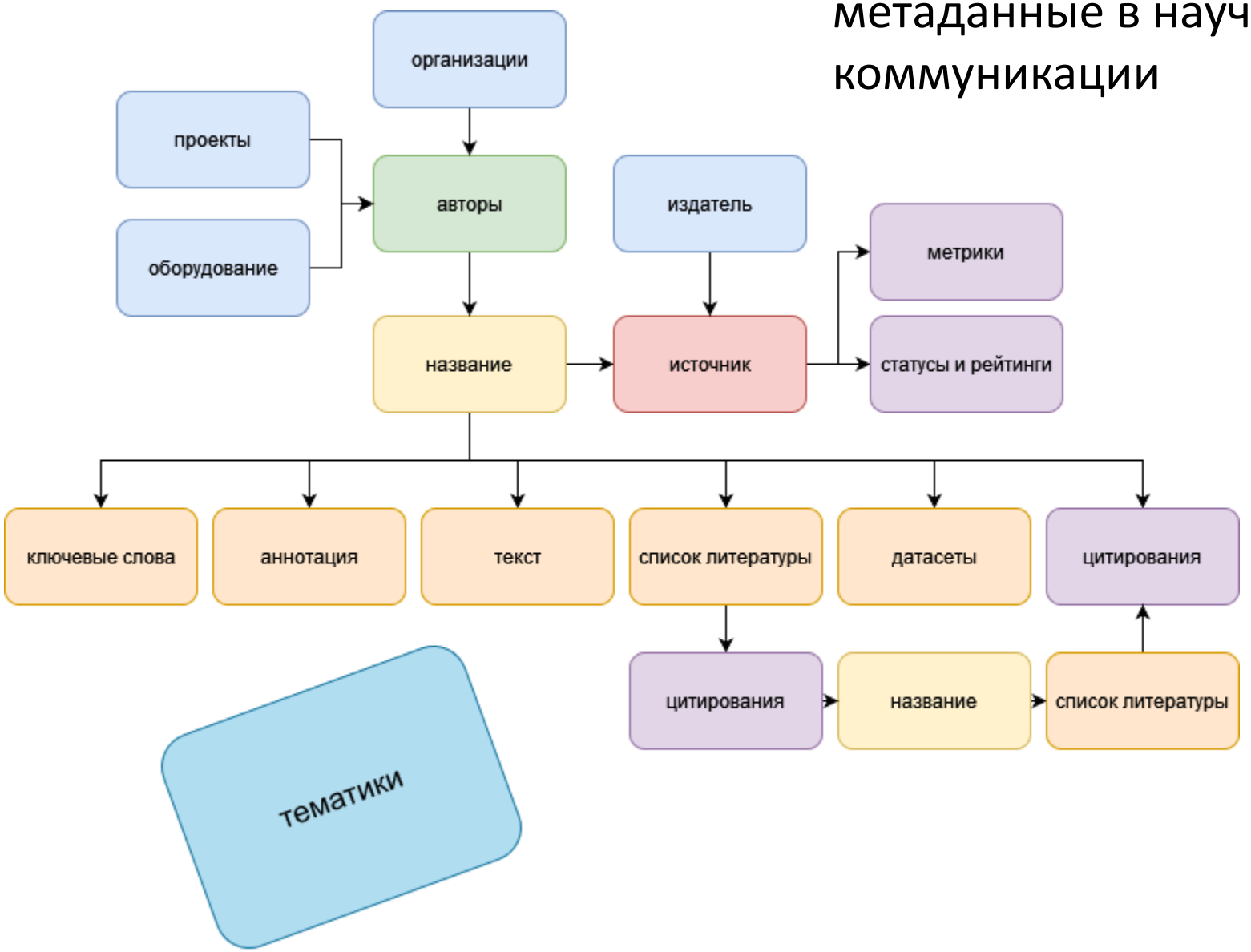
Ключевые тенденции

- Все открывается: и тексты, и данные, и метаданные
- Все стандартизируется и связывается

FAIR: data should be **findable**, **accessible**, **interoperable** and **re-usable**

- Авторов, данных и результатов все больше
- AI\ML все важнее
- Изменение системы оценки науки

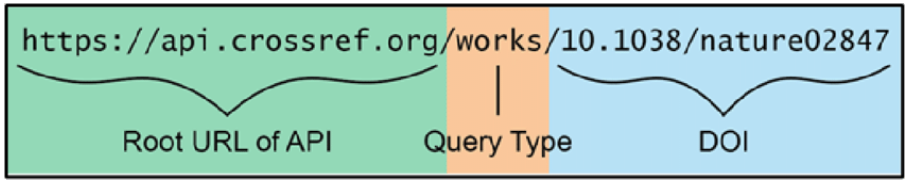
Основные данные и метаданные в научной коммуникации



Общенаучные базы и инструменты

- Метаданные публикаций по разным дисциплинам
- Все чаще открыты, бесплатны и доступны по API
- Покрытие отличается, но в основе большинства – Crossref и MAG + базы препринтов и открытых текстов
- Наборы метаданных очень разные, степень обработки – тоже, но много общих id
- В приоритете охват, а не селективность
- Проблемы с data curation (авторы, организации)
- Редко индексируют полные тексты
- Учет ссылок везде разный

Crossref



- Основа большинства других баз и систем обмена информацией о публикациях
- Данные от издателей в рамках присвоения DOI, разные издатели открывают разное
- Простой поиск: https://search.crossref.org/
- API бесплатный и хороший (есть платные опции): https://api.crossref.org/swagger-ui/index.html
- Очень много полей метаданных, но они часто пустые (в т.ч. аффилиации). Много аннотаций, но лицензии разные
- Нет полноценных профилей авторов и организаций
- Недоучет цитирований
- Нет данных от других doi-агентств (datacite)

Простое введение: https://www.crossref.org/documentation/retrieve-metadata/rest-api/a-non-technical-introduction-to-our-api/

Пример: https://api.crossref.org/works/10.1155/2014/413629

I4OC, OpenCitations, COCI

- **Initiative for Open Citations** (I4OC) – инициатива по принуждению издателей к размещению списков литературы в Crossref в открытом доступе

https://i4oc.org/#about

- **OpenCitations** – организация, продвигающая инфраструктуру отрытых ссылок
  - Ведет базу **COCI -** размеченный индекс Crossref doi-to-doi citations
  - https://opencitations.net/index/coci/api/v1
  - Фиксирует самоцитирования и возраст ссылок
  - Также ведет **OpenCitations Corpus**, агрегирующую информацию не только из Crossref

- Полностью открытая база-наследник **Microsoft Academic Graph**, пополняется из разных источников, в основном CrossRef
- Пока нет веб-интерфейса, работа через API и дампы
- https://docs.openalex.org
- https://sciguide.hse.ru/tech/api/ - коротко о работе с API на русском

https://www.semanticscholar.org/

- AI-powered research tool (TLDR)
- Сами собирают метаданные (робот+издательства)
- Полноценный веб-интерфейс
- Профили авторов
- Плохо с аффилиациями
- Нет выгрузки метаданных в веб-интерфейсе
- Делается одним из AI-лидеров (Allen Institute)
- Бесплатный API https://api.semanticscholar.org/api-docs/graph

LENS.ORG

https://www.lens.org/

- Делает небольшая австралийская компания Cambia
- Платно-бесплатная модель
- Объединяет статьи и патенты, большой охват
- Полноценная бесплатная версия
- Тестовый API (14 дней), много полей
- Нет полноценных авторских профилей
- Выгрузка метаданных из веб-интерфейса (без аффилиаций)

Как эти системы соотносятся по наличию метаданных для РФ?

https://podpiska.rfbr.ru/storage/reports2021/2022_meta_quality.html

Сравнение качества метаданных в БД CrossRef, Lens, OpenAlex, Scopus, Semantic Scholar, Web of Science Core Collection

Лутай А.В., Любушко Е.Э.

ФГБУ “Российский фонд фундаментальных исследований”

21.02.2022

Еще базы?

- **Dimensions**: от SpringerNature, дорогая, старается играть на рынке WoS\Scopus, скромная бесплатная версия: https://app.dimensions.ai/discover/publication
  - Лучше классификатор (на уровне публикаций), больше охват, есть патенты, клинические испытания, гранты, датасеты и т.д., есть экспертные списки
- **Wizdom.ai**: похоже на Dimensions, владельцы - Informa
- **Scilit.net**: еще одна база, от MDPI, бесплатная, ничего выдающегося.
- **FATCAT**: открытая база от Internet Archive, удобна для сбора информации о журналах
- **BASE**: огромный охват, много метаданных, либеральные лицензии, есть API https://www.base-search.net/

Отраслевые базы

Основной use case – поиск литературы профессионалами в отрасли

https://sciguide.hse.ru/sources/areas/

- Астрофизика: https://ui.adsabs.harvard.edu/
- Медицина и все вокруг: https://pubmed.ncbi.nlm.nih.gov/
- Физика высоких энергий: https://inspirehep.net/
- Компьютерные науки: https://dblp.org/

Publish or Perish

- Программа, работающая через API (большинство ресурсов) и веб-кроулинг (Google Scholar)
- Позволяет делать запросы, выгружать метаданные и разнообразные метрики (цитирования, Хирш etc.)
- https://harzing.com/resources/publish-or-perish/manual
- Источники данных: https://harzing.com/resources/publish-or-perish/manual/using/data-sources

Данные об источниках

- **ISSN Portal**: «слабое звено», платные данные. Бесплатно только таблица ISSN-L-to-ISSN
- **FATCAT**: https://fatcat.wiki/container/search
  - Набор метаданных невелик, но очень доступен https://search.fatcat.wiki/fatcat_container/_mapping
- **DOAJ:** много информации о журналах открытого доступа, хороший API https://doaj.org
- **Северные списки**:
  - Норвегия https://kanalregister.hkdir.no/publiseringskanaler/Forside?request_locale=en
  - Финляндия https://jfp.csc.fi/en/

Спасибо за внимание!

isterligov@hse.ru

scientometrics@hse.ru