Открытый воспроизводимый датасет публикационных метрик России

Иван Стерлигов, НИУ ВШЭ

isterligov@hse.ru

14 ноября 2023 г.

План выступления

- О чем это и зачем?
- Формат: скрипты, снэпшоты, публикации
- Технология
  - Источники
  - Методы
  - Воспроизводимость
- Первые результаты
  - Селективность
  - Направления
  - Международное сотрудничество
  - Издательства
  - Организации
- Планы на будущее

О чем это и зачем?

- Открытый и бесплатный,
- Регулярно обновляемый,
- Воспроизводимый

датасет ключевых библиометрических показателей России

- Науковедение
- Мониторинг развития науки
- Развитие открытой инфраструктуры открытой науки
- Развитие технологий работы с данными

Как это связано с…

- Официальной статистикой
- Всяческими рейтингами
- Национальными списками РФ
- Аналитическими продуктами НИУ ВШЭ
- Коммерческими поставщиками данных
- «Недружественными англосаксами»
- Гостехом и нацпроектами?

Никак.

Это научный некоммерческий проект.

Но ведь уже есть…

- «Индикаторы науки» ИСИЭЗ ВШЭ
- Сборники РИЭПП и других организаций
- Нелегальный доступ к старым-добрым SciVal и InCites
- Scimago Country Rank, Lens, Dimensions и т.п.

Этот проект отличают открытость, легальность, воспроизводимость, методология и сами наборы метрик.

Формат

- Датасет, описание и инструменты для его воспроизведения на Zenodo, all included
- Набор скриптов на Github
- Статья с описанием и анализом собранных данных (в работе)
- Интерактивные дэшборды (закрытый режим)

Источники данных

- Scopus через бесплатный общедоступный API
  - https://dev.elsevier.com/sc_apis.html
  - Хорошее покрытие, хорошие профили организаций, богатый инструментарий
  - Дефицит метаданных (аффилиации авторов)
  - Множество российских источников
  - Устаревшая классификация
  - Риски отключения
  - Юридические риски?
- OpenAlex через бесплатный общедоступный API (в работе)
  - https://openalex.org/
  - Нет рисков отключения, данные в public domain по CC-Zero
  - Разное покрытие (много лакун), больший охват
  - Разные профили организаций, гораздо хуже привязка к профилям организаций и странам
  - Потенциально богатый набор метаданных

Технологии сбора данных-1

Семейство скриптов S-pets

https://github.com/IvanSterligov/S-pets

Матричный принцип

	pubyear is 2018	pubyear is 2019	pubyear is 2020
affilcountry(Russia)	**X**		
affilcountry(Brazil)			
affilcountry(India)			

+ addon: doctype(ar or re or dp)

X = (affilcountry(Russia)) AND (pubyear is 2018) AND doctype(ar or re or dp)

Технологии сбора данных-2

- Применяем списки источников для фильтрации результатов
- Составлены из source-id (не ISSN)
- Включены в датасет
- Скрипты для работы со списками автоматизируют поисковые запросы (склейка и суммирование)

Технологии сбора данных-3

- Каждый индикатор = таблица
- Файл **tasks.csv** содержит все индикаторы, описания, запросы, наборы журналов и прочих фильтров, названия файлов с данными
- Его можно настраивать и модифицировать

indicator	script	main_input_file	source_list	addon	output_file
MIC1	count_matrix	mic1_main.csv	none	doctype(ar or re or dp or b	mic1_out.csv
MIC2a	count_matrix	mic1_main.csv	none	doctype(cp)	mic2a_out.csv
MIC3-2018	queryXid	query_countries.csv	lists_ni.csv	doctype(ar or re or dp) an	mic3_out_2018.csv
MIC3-2019	queryXid	query_countries.csv	lists_ni.csv	doctype(ar or re or dp) an	mic3_out_2019.csv
MIC3-2020	queryXid	query_countries.csv	lists_ni.csv	doctype(ar or re or dp) an	mic3_out_2020.csv

Технологии сбора данных-4

- Нужен API-ключ https://dev.elsevier.com/
- Скачайте весь датасет в любую папку
- Укажите путь к файлу с ключом и папке в скрипте **batch.py** и запустите его
- Если закончатся лимиты, уберите уже сделанное из **tasks.csv** и подождите неделю или укажите новый API-ключ
- Опционально запустите **combine_output.py** для объединения и повышения читаемости результатов

Используемые списки

- Scopus as is – справочно. Почему?
  - Очень разные по научному уровню журналы и прочие источники
  - Их набор постоянно меняется
- Nature Index
- Norway Publication Register
- Российские журналы в Scopus
- Топ-20 ведущих мировых издательств

Nature Index

Версия 2023 г.

“a reasonably consensual upper echelon of journals in the natural sciences” that “truly reflect the upper tiers of research achievement as judged by peers” https://www.nature.com/nature-index/faq

146 журналов

Life sciences	37
Physical sciences	19*
Earth sciences	11
Chemical sciences	15
Health sciences	64
Multidisciplinary	5

*Phys. Rev. B

берем целиком

Norwegian Publication Indicator

Наиболее известный национальный «белый список», составленный экспертами-предметниками по документированной и отработанной процедуре

https://npi.hkdir.no/ - об индикаторе

https://kanalregister.hkdir.no/publiseringskanaler/Forside.action?request_locale=en – сами списки

https://doi.org/10.2478/jdis-2018-0017 – научная статья архитектора системы

“The highest level is named “Level 2”. It includes only the leading and most selective international journals, series and book publishers ... The publication channels selected for Level 2 can only in total represent up to 20% of the world’s publications in each field”.

Используем уровень 2 и ограничиваемся журналами, потому что далеко не все журналы 1 уровня, книги и конференции есть в Scopus. Из журналов 2 уровня нет единичных по гуманитарным наукам.

2 уровня однозначной классификации, пока используем верхний (Health = 494, Natural = 606, Social = 458, Humanities = 538)

Российские журналы в Scopus

- Список на основе «полуофициального» elsevierscience.ru
- Обновлен с учетом данных Wikidata (РЦНИ), АНРИ (О.В. Кириллова) и вручную путем просмотра свежих пополнений Scopus
- Не включает журналы, потерявшие связь с Россией: RAMS, Las.Phys.Lett. и некоторые другие
- Включает 711 журналов
- Типология издателей: RAS, HEI, Russian business, Foreign business, non-RAS Institute, editorial office

Объекты анализа-1. Страны

RUSSIA

Для сравнения и исследований сотрудничества:

BRAZIL, CANADA, CHINA, FRANCE, GERMANY, INDIA, JAPAN, SOUTH KOREA, UNITED KINGDOM, UNITED STATES

Принцип: ведущие, разные. В целом соответствуют топу по числу публикаций

Объекты анализа-2. Организации

Вузы: МГУ, СПБГУ, УРФУ, НГУ, КФУ, ТГУ, МФТИ, ВШЭ, ИТМО, ПМГМУ, РНИМУ, СКОЛТЕХ

НИИ: ФТИ РАН, ФИАН, ИОХ РАН, ИЯФ СО РАН, ИПФ РАН, ИК СО РАН, ИОНХ РАН, ИБХ РАН, ИЦИГ СО РАН, МИАН, ПОМИ РАН

AF-ID берем из RIRO: https://openriro.github.io/

Для сравнения: Yale, Tsinghua, Humboldt

Принцип: ведущие, разные.

Некоторые результаты

- Динамика по Scopus, NI и NPI2
- Международное соавторство в разных форматах
- Перемены на уровне издательств
- Российские vs иностранные журналы
- Страны и организации

Прогноз изменения числа публикаций России: 2023 vs. 2021

На примере метрики общего числа публикаций Ar Re Dp в Scopus:

1. Вычисляем среднее соотношение 2023 и 2021 для остальных стран (кроме Индии и Китая) = «нормальный рост» = 81,1%
2. «Нормальное» число публикаций России = 2021*81,1% = 75650 публикаций в 2023.
3. Фактическое значение (58890) меньше на **16760** публикаций, или **22,2%**

Прогноз изменения числа публикаций России: 2023 vs. 2021

Nature Index:

Bio: -34%

Phys: -44%

Earth: -30%

Chem: -34%

Health: -32%

Multi: -40%

Totals: -39%

NPI2:

Natural\Engi Sci: -32%

Social Sci: -28%

Humanities: -29%

Health: -32%

Totals: -32%

Спасибо за внимание!

isterligov@hse.ru
ivan.sterligov@gmail.com
https://doi.org/10.5281/zenodo.10119095