Новая наукометрия 2:

Идентификаторы, API, интеграция

Иван Стерлигов

НИУ ВШЭ, 12 октября 2022 г.

isterligov@hse.ru

О чем этот семинар?

- **Recap**: новые источники данных и метаданных
- **Мечты и надежды**: Semantic Web, Open Knowledge Network, Linked Open Data, Resource Description Framework, Uniform Resource Identifiers
- **Реальность**: текущие инициативы государства, бизнеса и общества
- **Технологии**: идентификаторы как связующее звено
  - Публикации: DOI, ISBN, PMID, ArxivID etc.
  - Источники: ISSN и его виды
  - Авторы: ORCID и идентификаторы баз
  - Организации: ROR, GRID, Ringgold, российские данные
  - Деньги: номера грантов и Funder Registry
- **Основы работы с API**: прямой доступ, библиотеки, примеры (на python\requests)
- **Прозаическое**: организация учета и анализа публикаций сотрудников российского вуза

Recap: источники данных

- **Открытые и доступные**: лицензии, API, дампы
- **Инклюзивные**: в приоритете охват, берут препринты, патенты, датасеты etc
- **Связные**: основаны на общеупотребимых идентификаторах
- **Дырявые**: массовые пропуски и несоответствия, особенно по авторам и аффилиациям
- **Сырые**: проблемы с data curation и профилями
- В основном основаны на CrossRef и MAG

Big idea: всеобъемлющий общедоступный граф знаний…

…делающий знания людей понятными машинам.

Все хорошее для всех, в том числе:

- Автоматизация, коммуникация и обогащение исследований
  - Сервисы для ученых и разработчиков
  - Замена ученых
- Мониторинг, аналитика, оценка, распределение ресурсов: ученые, чиновники, фонды
- Построение\наполнение локальных информационных систем: вузы, журналы, библиотеки

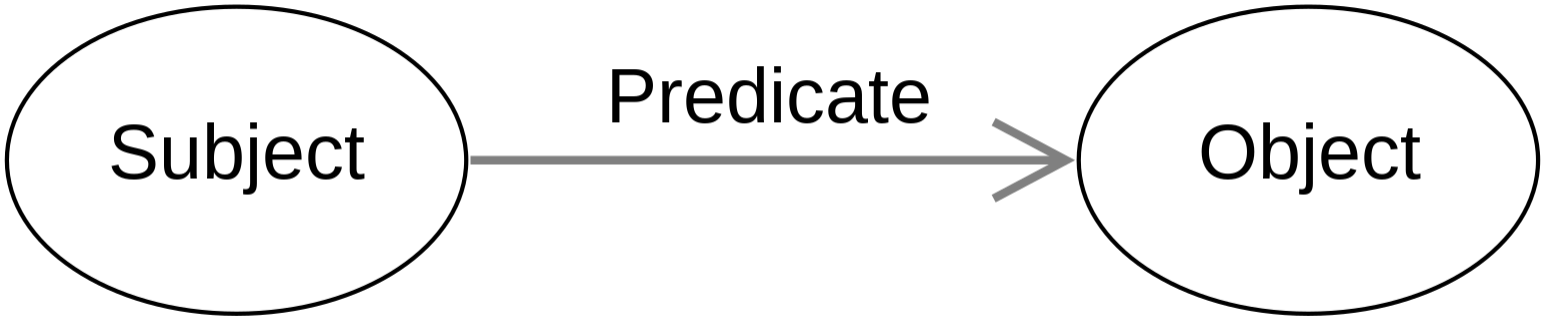
Semantic Web

«I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize».

Tim Bernes-Lee, 1999, ISBN: 978-0-06-251587-2.

Resource Description Framework

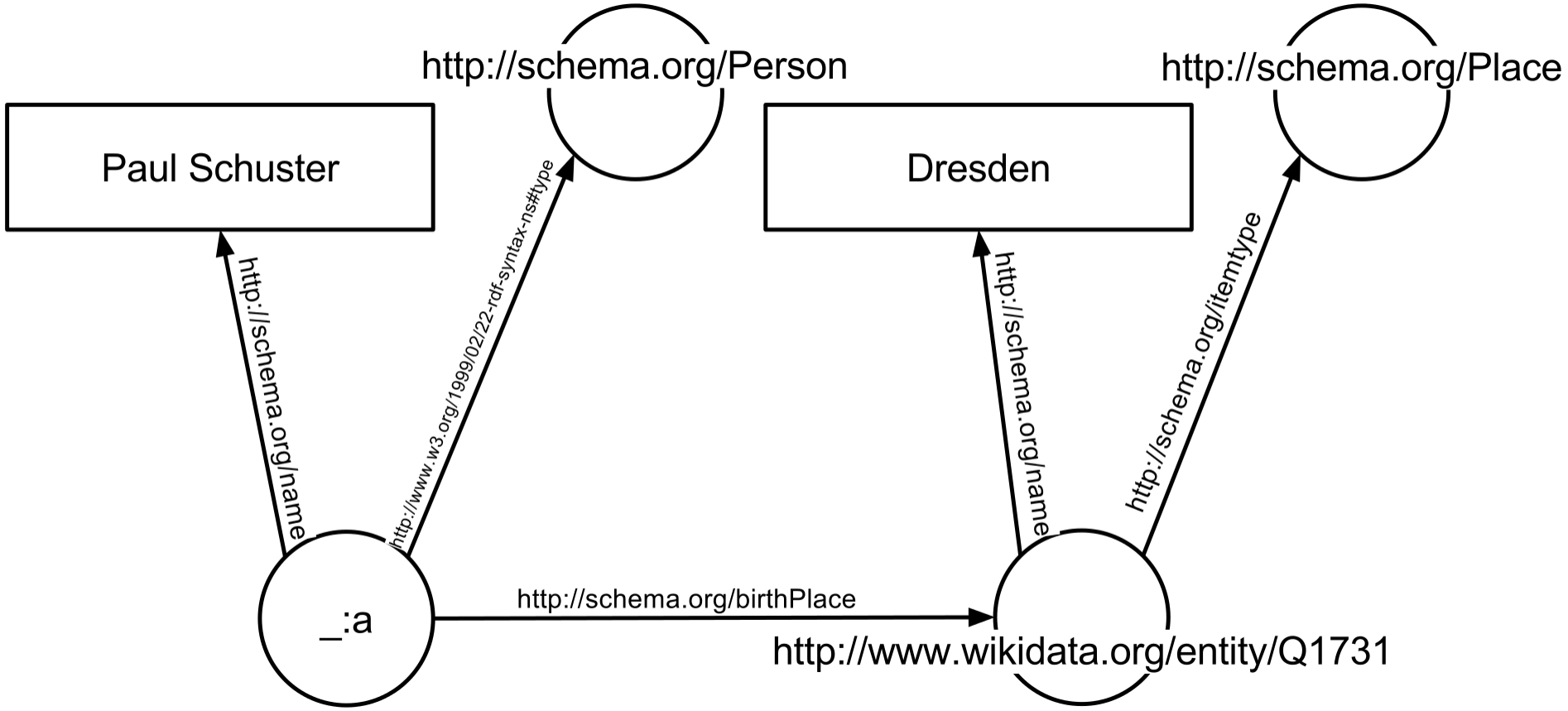
- Магистральный подход к описанию данных для графов знаний
- Построен на тройках:



- У каждого компонента триады – уникальный ID (в случае Semantic Web это обязательно ссылка)
- Свой язык запросов: SPARQL

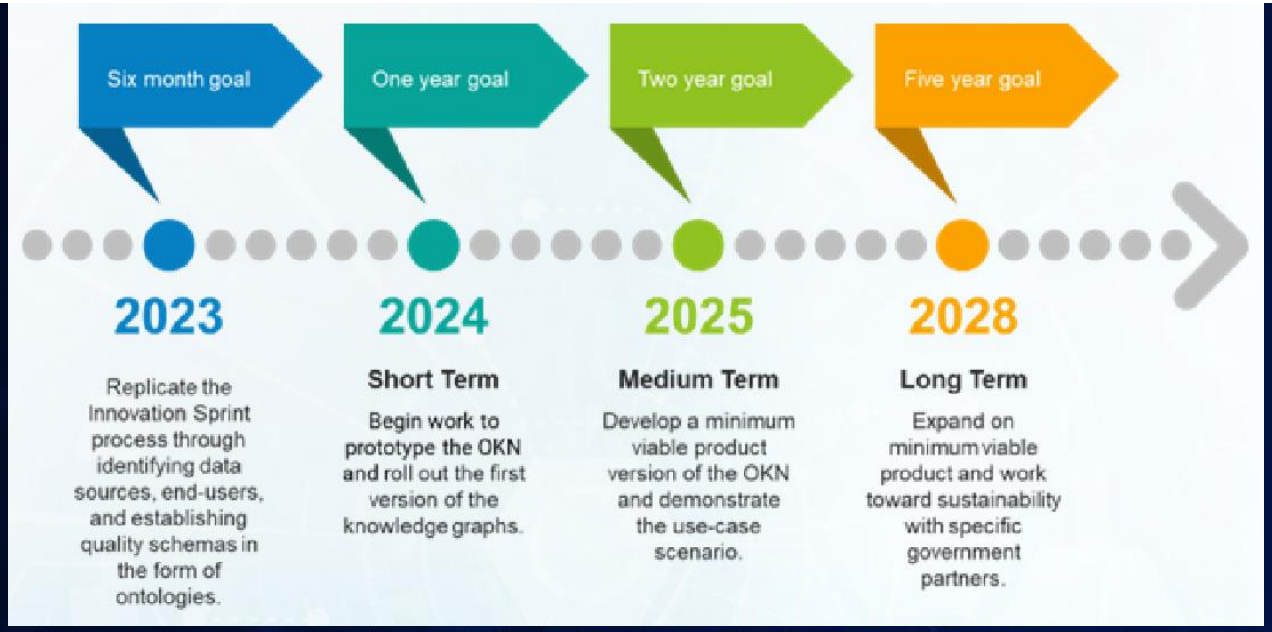
Semantic Web

```
<div vocab="https://schema.org/" typeof="Person">
  <span property="name">Paul Schuster</span> was born in
  <span property="birthPlace" typeof="Place"
href="https://www.wikidata.org/entity/Q1731">
    <span property="name">Dresden</span>.
  </span>
</div>
```



Официальный сегмент: OKN Roadmap

https://nsf-gov-resources.nsf.gov/2022-09/OKN%20Roadmap%20-%20Report_v03.pdf



OPEN KNOWLEDGE NETWORK ROADMAP:

POWERING THE NEXT DATA REVOLUTION

SEPTEMBER 2022

Коммерческий (и главный на сегодня)

Google Knowledge Graph

совсем не open

https://www.google.com/search?q=hse+university

500 billion facts on 5 billion entities (2020)

А также:

Microsoft

Yandex

LinkedIn

Apple

Amazon и т.д.

Некоммерческий и открытый

Item	Property	Value
Q42	P69	Q691283
Douglas Adams	educated at	St John's College

https://www.wikidata.org

Введение: https://www.wikidata.org/wiki/Help:About_data

Визуализация графа: https://angryloki.github.io/wikidata-graph-builder

Кроме того:

DBpedia (данные, извлеченные из Википедии)

Научная часть Wikidata

SCHOLIA

https://scholia.toolforge.org/

Пример профиля автора:
https://scholia.toolforge.org/author/Q26322

- Мало данных (только Wikidata)
- Издержки вики-подхода (messy\inaccurate data)

От великого до малого?

FAIR, Linked Data, использование глобальных идентификаторов и открытых API

Подходы, инструменты и данные «мегаграфов» можно и нужно использовать при создании и развитии малых систем, например:

- Проект
- Университет \ НИИ
- Грантовый фонд
- Издательство
- Министерство

Процесс может быть двусторонним

Связующий элемент - URI

Uniform Resource Identifier is a unique sequence of characters that identifies a logical or physical resource used by web technologies

Подвиды:

Uniform Resource Locator (является ссылкой)

Uniform Resource Name (не является ссылкой)

Близкая вещь:

Handle System

Связующая технология - API

Application Programming Interface – интерфейс взаимодействия программ

Множество стандартов и подходов. Основные понятия:

- Private\Public, Keys
- Endpoints
- Calls, Rates, Requests, Responses
- HTTP, GET, Header, Body
- Output formats (JSON, XML)
- Singletons \ Collections, cursor\paging

Рекомендуем Python и Requests https://requests.readthedocs.io/en/latest/

Хотите качать быстрее? https://docs.aiohttp.org/en/stable/

Ничего не знаете про Python? https://www.w3schools.com/python/

```json
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [
      "Catherine",
      "Thomas",
      "Trevor"
  ],
  "spouse": null
}
```

Основной формат: **JSON** (JavaScript Object Notation)

Если без экселя совсем никак...

```
import requests
import pandas as pd
from flatten_json import flatten


url=r'https://inspirehep.net/api/institutions?q=addresses.country_code:RU&size=1000&page=1'
rawjson=requests.get(url=url).json()['hits']['hits']
dic_flattened = [flatten(d) for d in rawjson]
df = pd.DataFrame(dic_flattened)
df.to_excel(r'c:\out.xls')
```

Ключевые идентификаторы в научной коммуникации

	глобальные	**специальные**
Публикации	DOI, ISBN	EID, UT, EDN
Источники	ISSN	SourceID
Авторы	ORCID	SPIN, ResearcherID
Организации	ROR	Affiliation ID
Деньги	Funder Registry?	РНФ
Тематики и концепты	OECD FoS? FoR? MAG? Wikidata?	ASJC

Публикации. DOI

Ресолвер Префикс Суффикс

- https://doi.org/10.1000/182
- Создан в 90-е американским издательским бизнесом, сейчас применяется почти всеми, в основе - Handle
- Может присваиваться любым объектам, в т.ч. статьям, книгам, главам в них, датасетам, препринтам, журналам. У разных версий объекта – обычно разные DOI
- Выдается («чеканится», minting) агентствами-регистраторами
- Главное и рекомендуемое – **CrossRef** (из-за сбора и предоставления метаданных из своей БД). Для данных и «серой литературы» часто используется **DataCite**
- Ресолвится в URL. Если не ресолвится – «битый»

API: https://api.crossref.org/swagger-ui/index.html

Пример по статье: https://api.crossref.org/v1/works/10.1088/0004-637X/722/2/971

Также можно использовать основанные на Crossref системы, например, OpenAlex

Введение и простые примеры:

https://sciguide.hse.ru/tech/api/

Публикации: ISBN

https://www.isbn-international.org

- Древний, не ресолвится, содержит только цифры и дефисы, разное число символов
- Свой у каждого издания книги (в т.ч. у разных форматов, в т.ч. электронных)
- Выдается национальными агентствами

https://www.bookchamber.ru/index.html

Обзор API: https://www.vinzius.com/post/free-and-paid-api-isbn/

Постепенно замещается DOI

Публикации: прочие важные ID

- ArxivID https://arxiv.org/help/api/basics
- PMID https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html

Серийные источники: ISSN

- Древний, выдается национальными регистраторами
- Печатная и электронная версии имеют разные ISSN
- «Основной» - ISSN-L
- https://www.issn.org/ru/services-et-prestations/services-en-ligne/acces-a-la-table-issn-l/
- Официальный API платный. Можно попробовать:

https://doaj.org/api/docs
https://guide.fatcat.wiki/entity_container.html

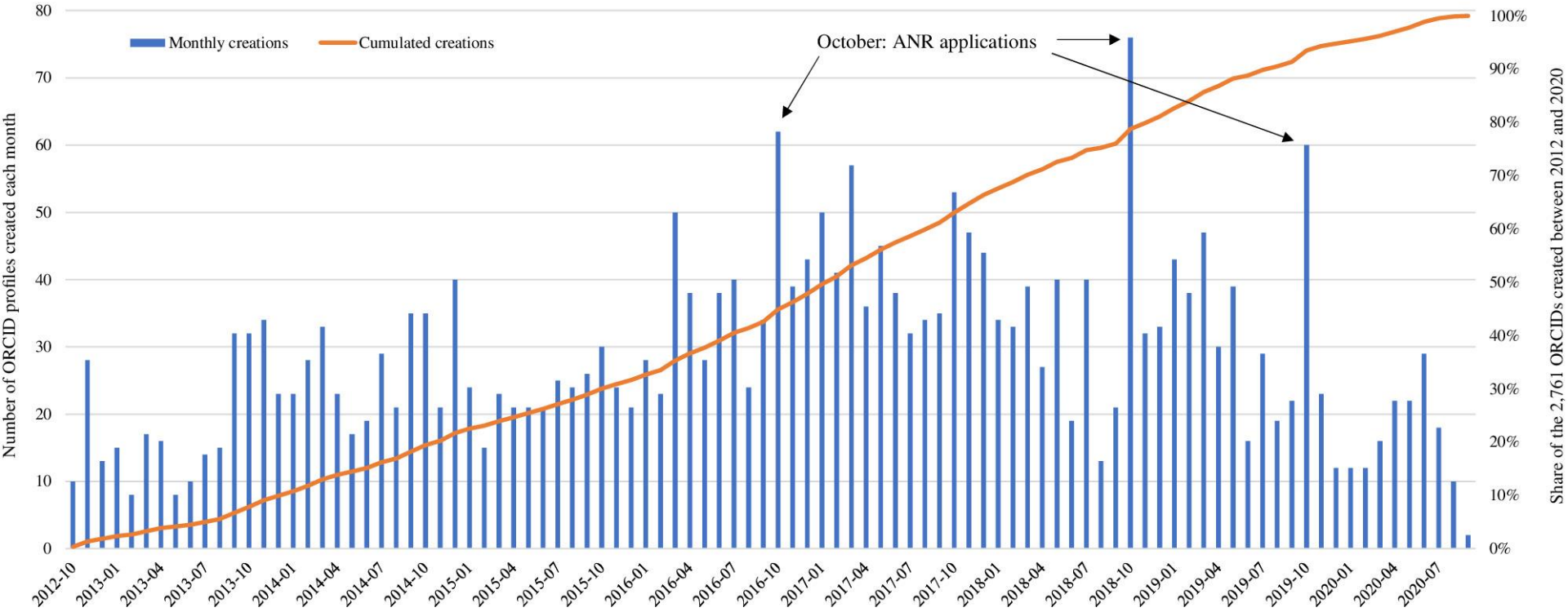
https://v2.sherpa.ac.uk/api/

Авторы: ORCID

- https://orcid.org/0000-0002-1825-0097
- Выдано порядка 15 миллионов ORCIDs
- Профили либо публичные, либо скрытые (требуется разрешение автора)
- Потенциально содержат массу информации: карьера, публикации, проекты, рецензирование и т.д.
- Многовато глюков и мусора
- Есть API как на чтение (общедоступный), так и на запись (для членов ORCID)

https://info.orcid.org/documentation/api-tutorials

Авторы: внедрение ORCID



https://doi.org/10.1002/leap.1451

Статистика по странам:
https://doi.org/10.3389%2Ffrma.2022.779097
Проникновение метаданных:
https://api.crossref.org/v1/works?filter=has-orcid:true

Организации: ROR

- GRID был до 2021 г., теперь только ROR
- Research Organization Registry https://ror.org/

https://ror.org/055f7t516

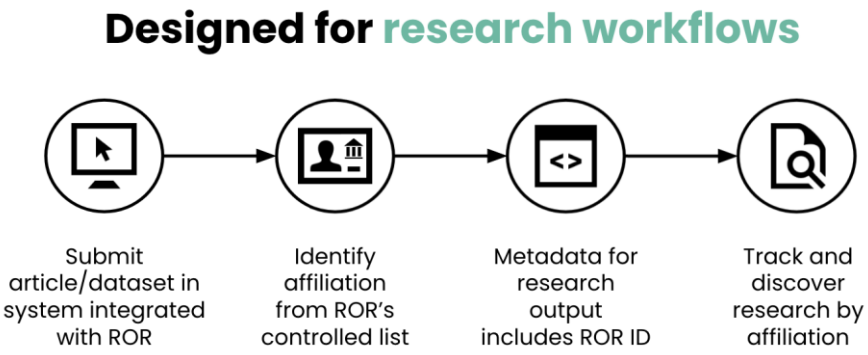
Любой может запросить дополнение, изменение, слияние, удаление, валидирует сообщество

- Интеграция с CrossRef (DOI) и ORCID
- API https://ror.readme.io/docs/rest-api (ограничения ElasticSearch)

Коммерческий конкурент: Ringgold

ROR

Как это работает?



Датасет-переходник, объединяющий идентификаторы российских организаций

https://openriro.github.io/

Финансирование

Funder Registry – проект CrossRef (подарен миру Эльзевиром), открытый, CC0

API: https://api.crossref.org/v1/works?filter=funder:10.13039/100000001

Идентификатор – DOI:

http://data.crossref.org/fundingdata/funder/10.13039/100000001

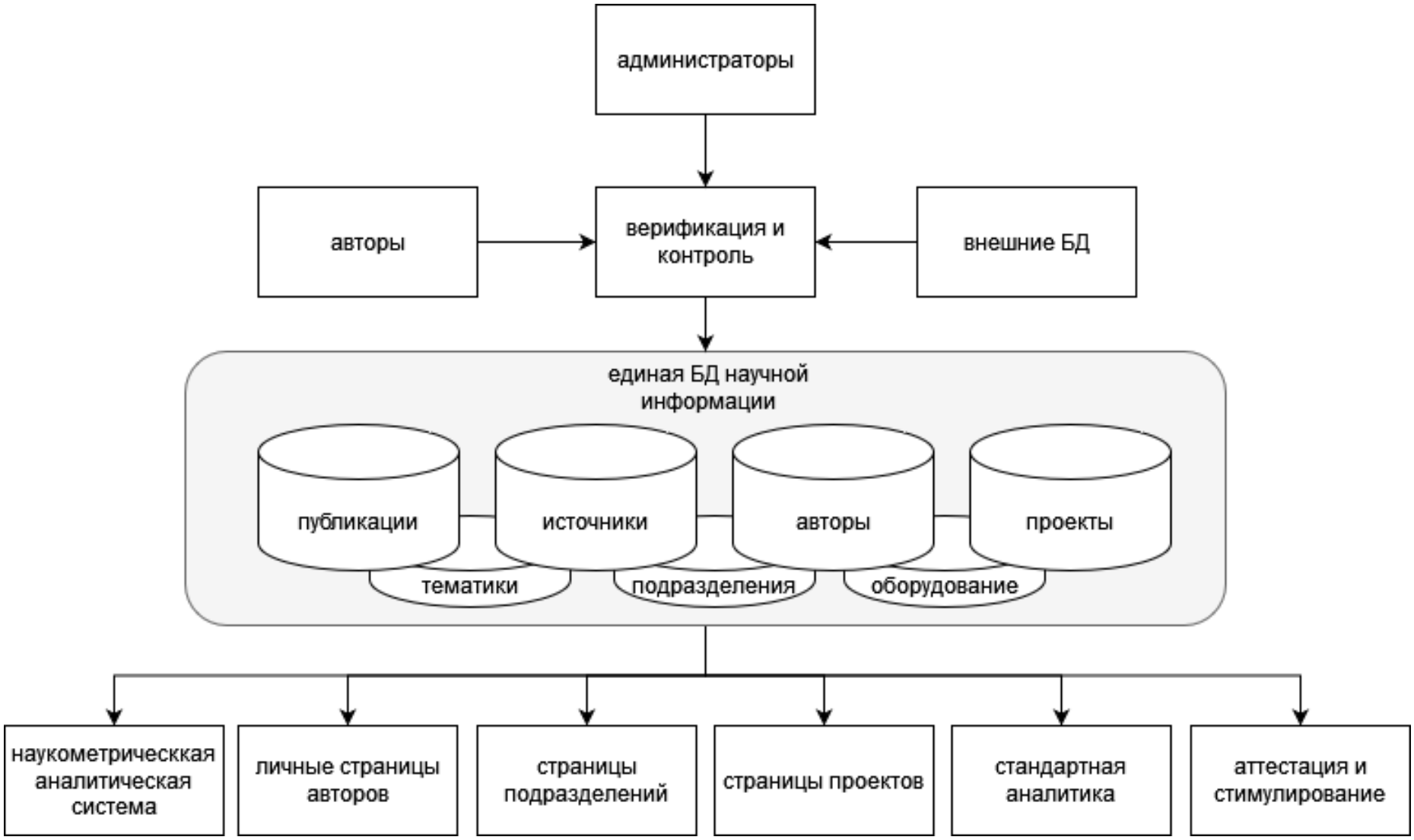
Ожидаем существенный рост из-за Nelson Memo

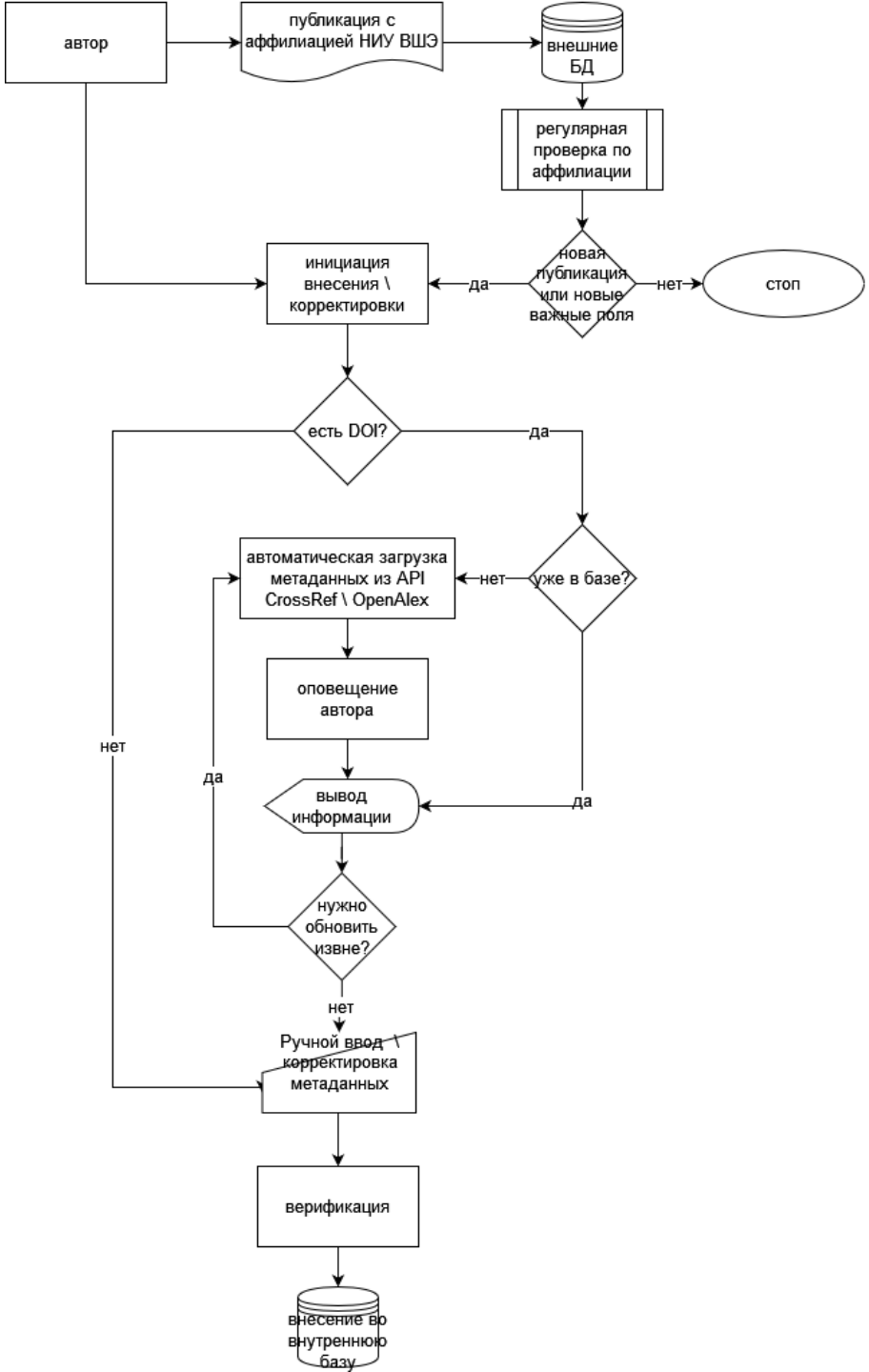
Зачем все это?

- Наполнение внутренних баз
- Обогащение внешних баз
- Построение мэшапов
- Интеграция и data wrangling
- Облегчение жизни авторов: автозаполнение полей, выбор из списков
- Аналитика
- Исследования

Например:

https://doi.org/10.3389%2Ffrma.2022.835139


администраторы
авторы
верификация и контроль
внешние БД
единая БД научной информации
публикации
источники
авторы
проекты
тематики
подразделения
оборудование
наукометрическая аналитическая система
личные страницы авторов
страницы подразделений
страницы проектов
стандартная аналитика
аттестация и стимулирование


автор
публикация с аффилиацией НИУ ВШЭ
внешние БД
регулярная проверка по аффилиации
новая публикация или новые важные поля
да
нет
стоп
инициация внесения \ корректировки
есть DOI?
да
нет
уже в базе?
нет
автоматическая загрузка метаданных из API CrossRef \ OpenAlex
оповещение автора
да
вывод информации
да
нужно обновить извне?
нет
Ручной ввод \ корректировка метаданных
верификация
внесение во внутреннюю базу

Спасибо за внимание!

isterligov@hse.ru