OpenAlex

Vladimir Batagelj
IMFM, UP IAM

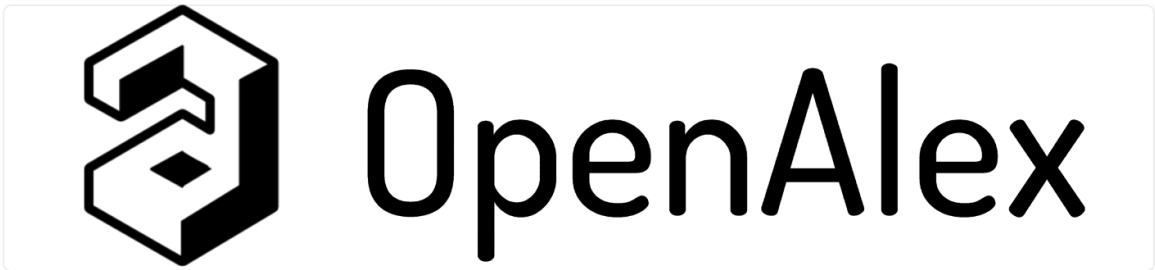
Ponedeljkov seminar računalništva in informatike
Koper, 2. september 2024

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (August 31, 2024 at 16 : 23): slides PDF
https://github.com/bavla/OpenAlex

OpenAlex

Overview



OpenAlex is a fully open catalog of the global research system. It's named after the ancient Library of Alexandria and made by the nonprofit OurResearch.

This is the **technical documentation for OpenAlex,** including the **OpenAlex API** and the **data snapshot.** Here, you can learn how to set up your code to access OpenAlex's data. If you want to explore the data as a human, you may be more interested in **OpenAlex Web.**

Data

The OpenAlex dataset describes scholarly *entities* and how those entities are connected to each other. Types of entities include works, authors, sources, institutions, topics, publishers, and funders.

Together, these make a huge web (or more technically, heterogeneous directed graph) of hundreds of millions of entities and billions of connections between them all.

OpenAlex

OpenAlex is a fully open catalog of the global research system [1]. It's named after the ancient Library of Alexandria and made by the nonprofit OurResearch.

OpenAlex launched in January 2022 with a free API and data snapshot. It is considered an alternative to the Microsoft Academic Graph (MAG), which retired on Dec 31, 2021 [2].

French Ministry of Higher Education and Research partners with OpenAlex to develop a fully open bibliographic tool.
The CNRS has unsubscribed from the Scopus publications database.
Wikipédia

OpenAlex

Blog

Researchers, funders, and organizations around the world rely on scientific knowledge graphs to find, perform, and manage their research. For decades, only paywalled proprietary systems have provided this information and they have become unaffordable (costing libraries $1B annually); uninclusive (systematically excluding works from some fields and geographies); and unavailable (even paid subscribers are limited in their use of the data).

OpenAlex indexes more than twice as many scholarly works as the leading proprietary products and the entirety of the knowledge graph and its source code are openly licensed and freely available through data snapshots, an easy to use API, and a nascent user interface.

OurResearch has a decade of sustained experience developing tools that advance open science. Funds from Arcadia will fuel the development needed to establish OpenAlex as the go-to scientific knowledge graph for researchers and organizations around the world. Long-term sustainability of OpenAlex will be achieved through value-add premium services.

OpenAlex

History

OpenAlex History

May 2021- Microsoft announced MAG sunsetting
Dec 2021- MAG discontinued
Jan 2022- OpenAlex beta launched
May 2022- User Group launched
August 2022- Full text search
December 2022- Customer support ticket system
March 2023- Premium offering launched
July 2023- Improved author disambiguation launched

Webinar: Introducing OpenAlex 10.30, 18.30

OpenAlex

Comments

OpenAlex solves several important questions for the analysis of bibliographic data:

1. identification of bibliographic units (IDs, disambiguation)
2. free access (share derived data, Download to your machine)
3. improving content through user participation (Submit a request)

We are working on a project of higher-level bibliographic services using bibliographic data analysis to advise the user. For example: the selection of reviewers, the selection of a journal to publish an article, etc.
A good example is the OpenAlex report of bibliographic data for an individual unit. For example, an individual author. To display our bibliography, we include a link to our website. Photo!?

https://openalex.org/authors/A5001676164

OpenAlex

Comments

As a data analyst, I miss short names for individual units (Garfield, journal abbreviations, etc.).

Person names are not structured (First, Mid, Last).

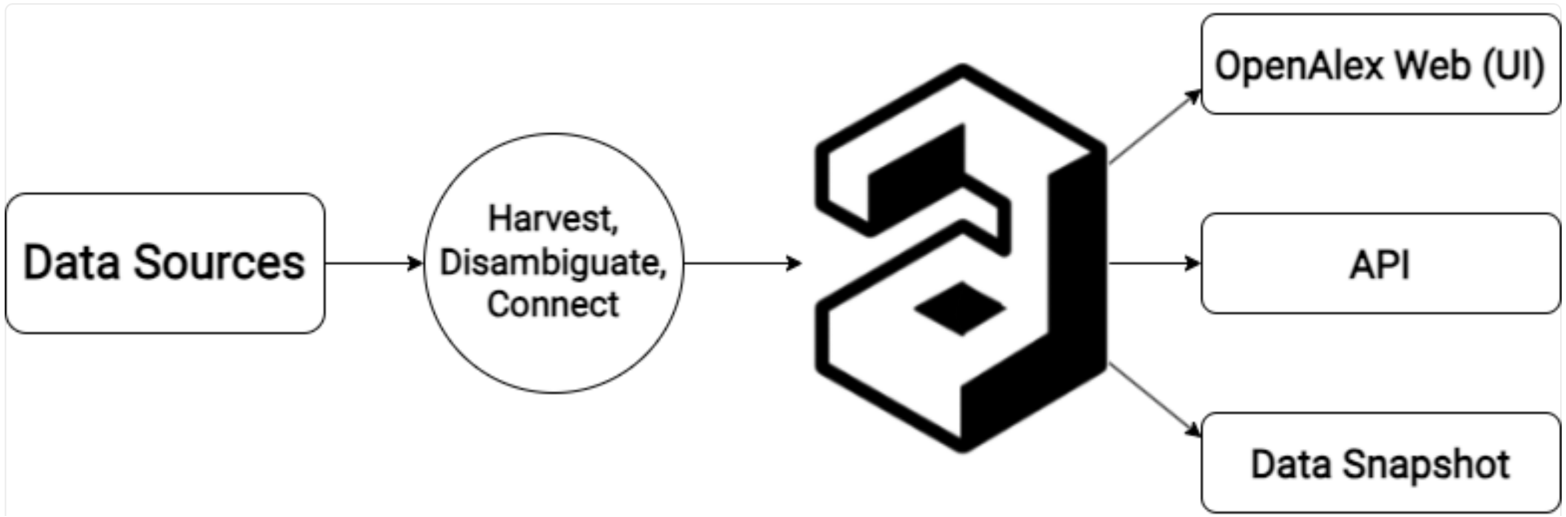
The problem of author countries – my example of an "extinct" country. W2033820728, W2059649701
(JSONview) Click API, see institutions

Missing relations to derived works (preprint – published, translation, book edition, etc.).

To ensure the OpenAlex longevity – UNESCO?

OpenAlex

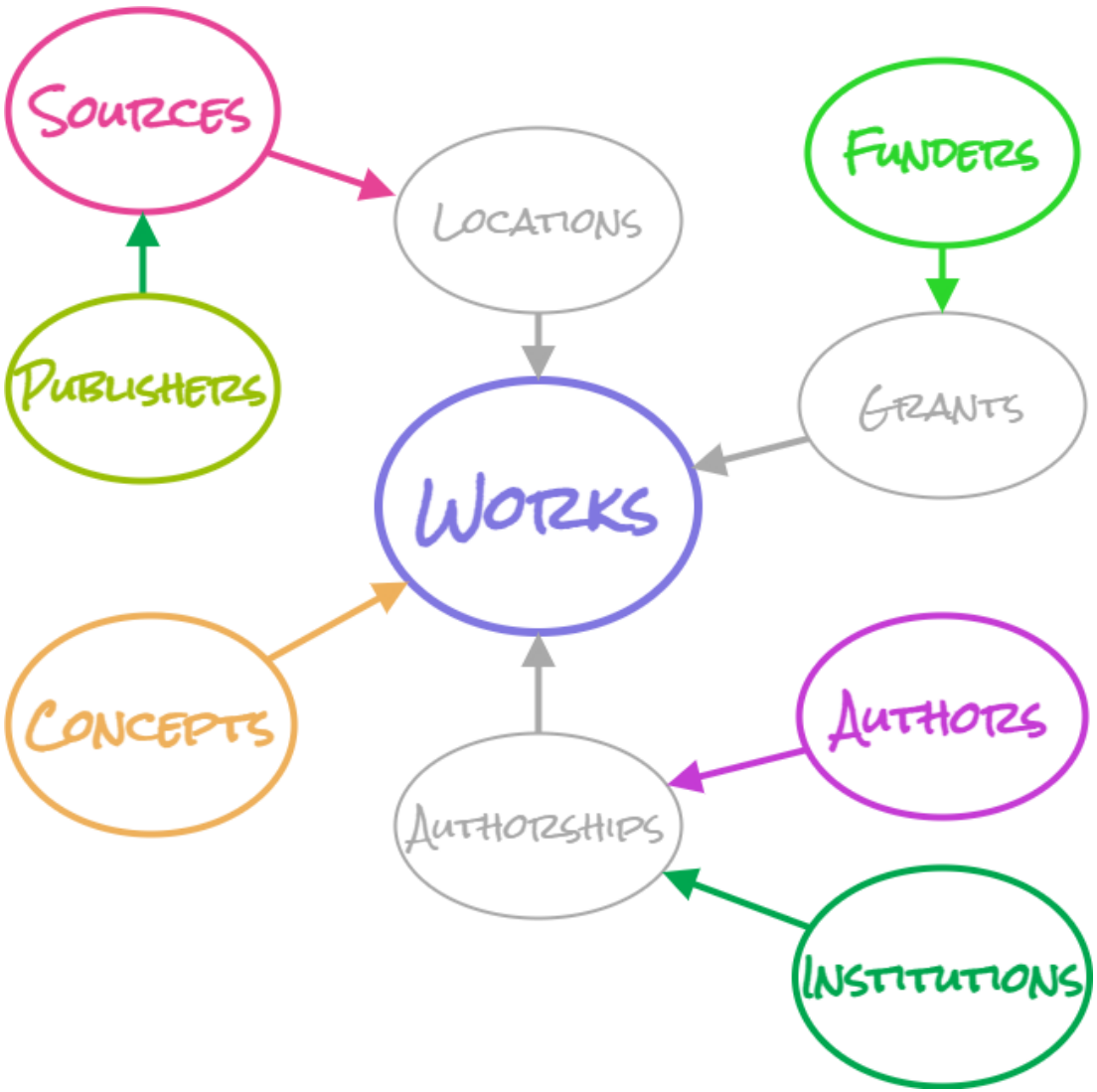
How it works



OpenAlex is based on 7 types of units (entities): **W**(ork), **A**(uthor), **S**(ource), **I**(nstitution), **C**(oncept), **P**(ublisher), or **F**(under)

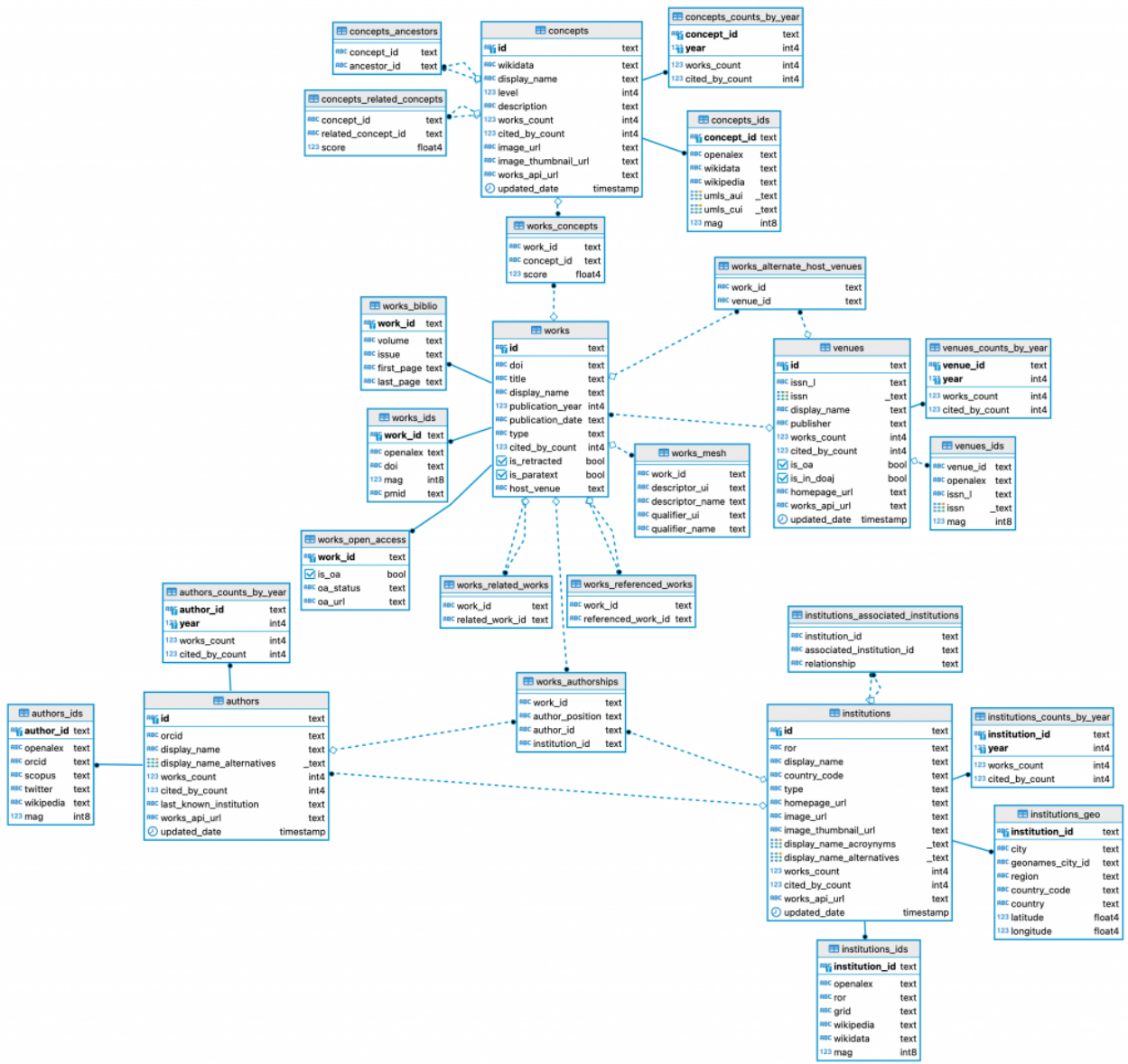
OpenAlex

Types of bibliographic units



OpenAlex

Scheme



OpenAlex

Using web browser

- OpenAlex site https://openalex.org/
- Known author ID https://openalex.org/A5001676164
- Work with DOI https://api.openalex.org/works/https://doi.org/10.1007/s11192-012-0940-1
- Known work ID https://openalex.org/W2083084326
- Name of the institution https://api.openalex.org/institutions?search=imfm
- Known institution ID https://openalex.org/institutions/I4210106342

R
Some functions

OpenAlex

Using API from program

The OpenAlex API is available at https://api.openalex.org. Its response is returned in JSON format. Here is an R code using the OpenAlex API

```
setwd(wdir <- "C:/work/OpenAlex/API")
library(httr); library(jsonlite)
wd <- GET("https://api.openalex.org/works",
  query = list(
    search="handball",
    filter="publication_year:2015",
    select="id,title",
    page="2", per_page="200"))
names(wd)
wc <- fromJSON(rawToChar(wd$content))
names(wc); names(wc$meta)
wc$meta$count; str(wc$results)
```

It returns the second page (with up to 200 entries) on works on handball published in the year 2015. Only information about works ID and title is returned.

OpenAlex

Search, filter, select

Using **search** we can search for a given search text across titles, abstracts, and full-text. Using a **filter** we can limit our search to units satisfying given conditions. Using **select** we can select data fields that will appear in results.

List of work IDs with titles

The OpenAlex API uses paging – the list data are provided by pages. The **basic paging** (up to 10 000 units) is based on two parameters page and `per_page`). **Cursor paging** is a bit more complicated than basic paging, but it allows us to access as many records as we like.

We developed an R package `OpenAlex2Pajek` for constructing a collection of Pajek bibliographic networks on selected topic from OpenAlex.

Conclusions

1. Open-houses
2. Webinars
3. Google user group
4. GitHub/topic/OpenAlex
5. Applications
    1. Webinar: How EPFL uses OpenAlex for tailor-made scientometrics and benchmarking between Universities
    2. OpenAlex Scholar in Emacs
6. Delgado-Quirós, L; Ortega, JL: Completeness degree of publication metadata in eight free-access scholarly databases [3]
7. [4], [5]

Acknowledgments

The computational work reported in this paper was performed using a collection of R functions `OpenAlex`, R program `OpenAlex2Pajek`, and the program Pajek for analysis of large networks. The code and data are available at Github/Bavla/OpenAlex.

This work is supported in part by the Slovenian Research Agency (research program P1-0294, research program CogniCom (0013103) at the University of Primorska, and research projects J5-2557, J1-2481, and J5-4596), and prepared within the framework of the COST action CA21163 (HiTEc).

References I

1. Priem, J., Piwowar, H. & Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* (2022).
2. Chawla, D. S. Massive open index of scholarly papers launches. *Nature* (2022).
3. Delgado-Quirós, L. & Ortega, J. L. Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 1–36 (2024).
4. Zhang, L., Cao, Z., Shang, Y., Sivertsen, G. & Huang, Y. Missing institutions in OpenAlex: possible reasons, implications, and solutions. *Scientometrics*, 1–23 (2024).
5. Jiao, C., Li, K. & Fang, Z. How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Scientific Data* **10**, 737 (2023).