OpenAlex2Pajek

an R-library for creating bibliometric networks

Vladimir Batagelj
IMFM, UP IAM

Applied statistics
Koper, 22-24. september 2024

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (September 16, 2024 at 04:18): slides PDF
https://github.com/bavla/OpenAlex

OpenAlex

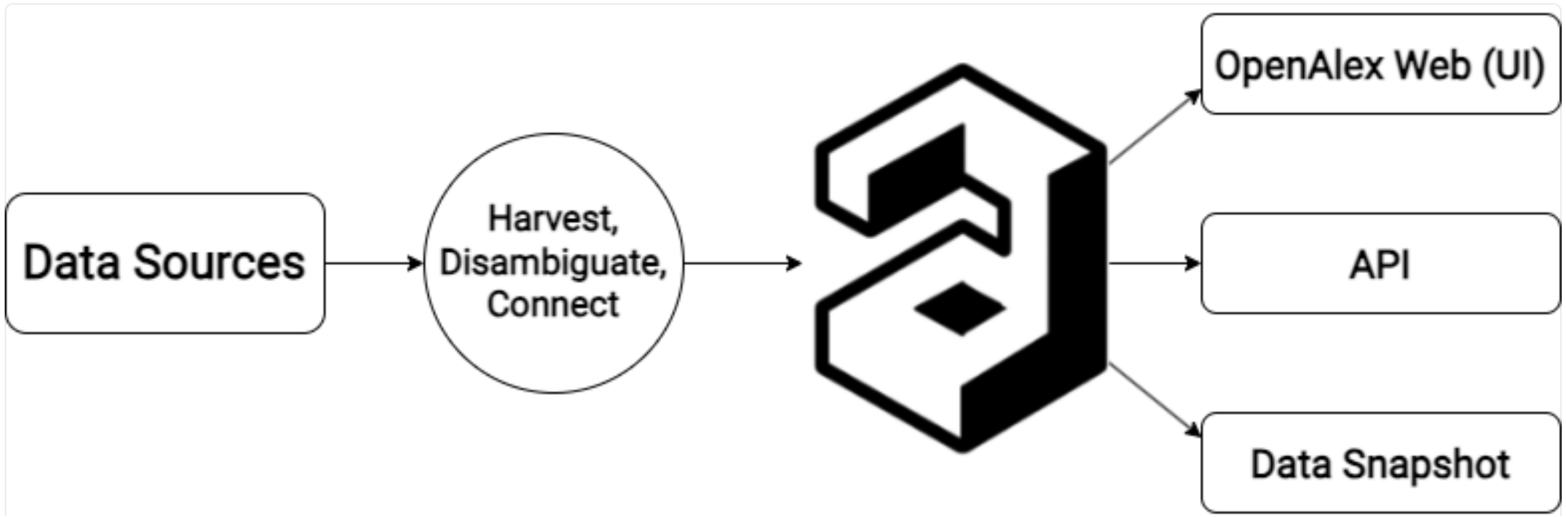
OpenAlex is a fully open catalog of the global research system [1]. It's named after the ancient Library of Alexandria and made by the nonprofit OurResearch.

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

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.

OpenAlex

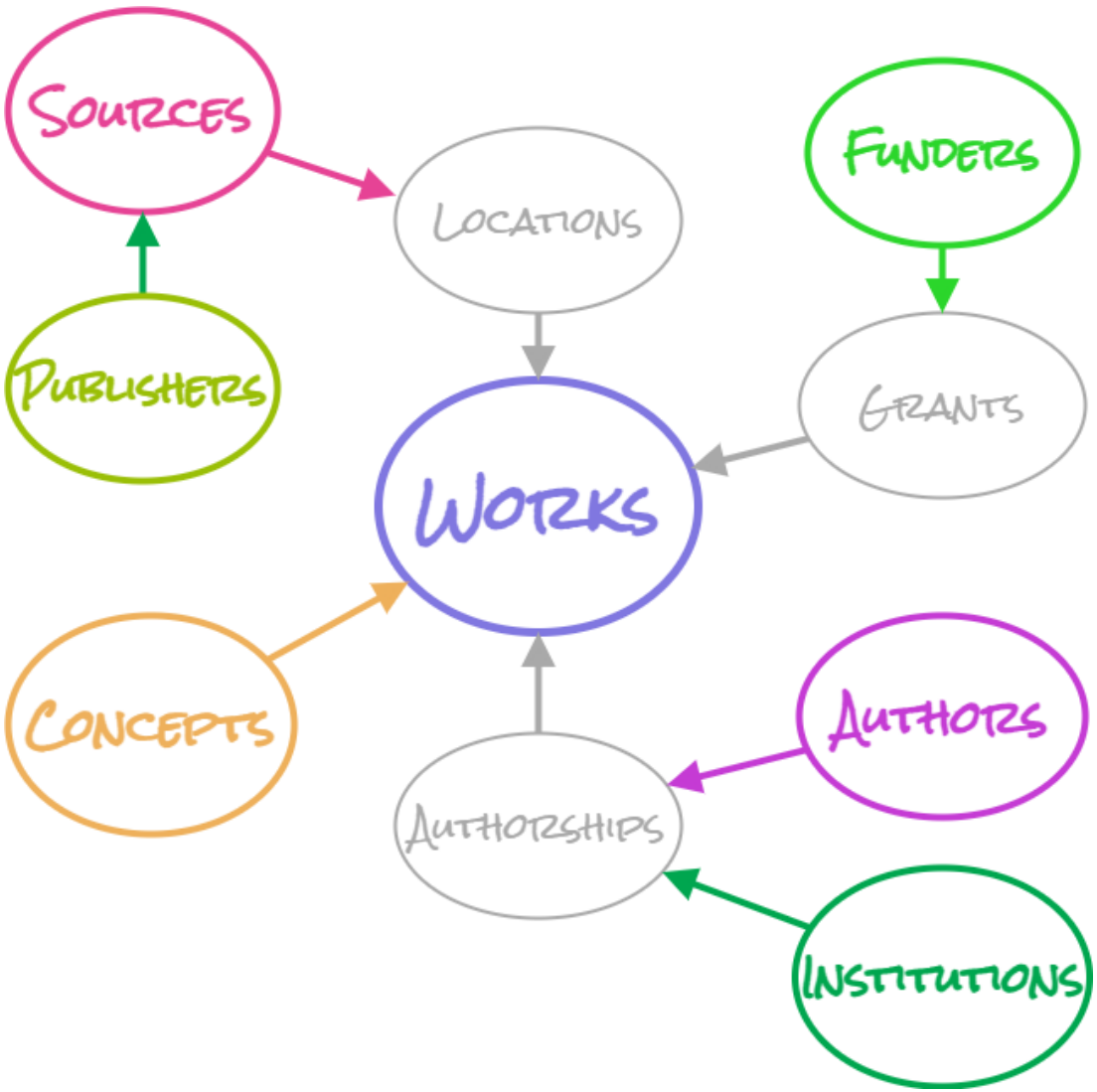
How it works



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

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.

https://openalex.org/authors/A5001676164

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

OpenAlex

Using API from a 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.

R – max number of authors; Some functions

OpenAlex2Pajek

We developed an R package `OpenAlex2Pajek` for constructing a collection of Pajek bibliographic networks on selected topics from OpenAlex. Currently, `OpenAlex2Pajek` contains three main functions `OpenAlex2PajekCite`, `OpenAlex2PajekAll`, and `coAuthorship`.

We split the process of creating the collection of bibliographic networks into two parts:

- determining the set W of relevant works using the **saturation approach** (Batagelj et al., 2017, page 506],
- creation of the network collection for the works from W.

The set W is determined iteratively using the function `OpenAlex2PajekCite` and the collection is finally created using the function `OpenAlex2PajekAll`.

. . . OpenAlex2Pajek

The collection contains the citation network **Cite** and two-mode networks: authorship **WA**, sources **WJ**, keywords **WK**, countries **WC**, and work properties: **publication year**, **type** of publication, the **language** of publication, **cited by** count, **countries distinct** count, and **referenced** works.

OpenAlex

Selected Institution

OpenAlex

Co-authorship

From OpenAlex we can collect the data about the co-authorship between world countries. To get a selected country, for example, SI, collaboration list we use the query

https://api.openalex.org/works?filter=authorships.countries:SI&group-by=authorships.countries

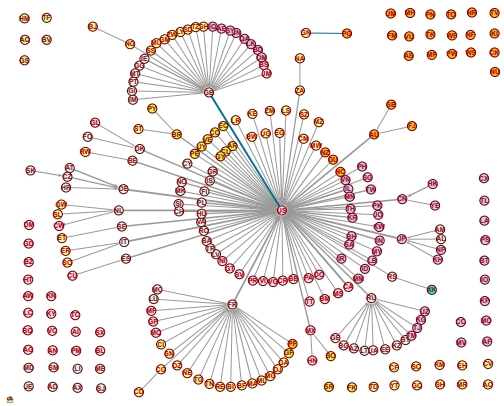
We developed a function `coAuthorship` that creates a temporal network describing the co-authorship between world countries in selected time intervals.

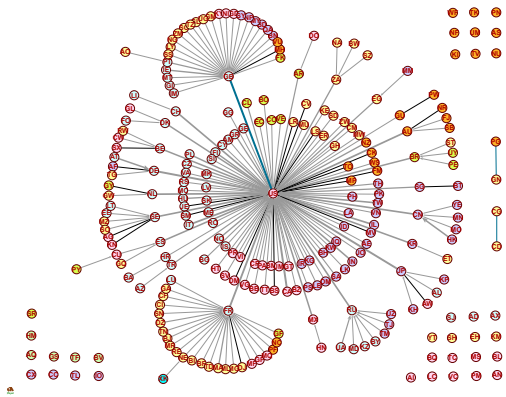
OpenAlex is using the ISO 3166-1 alpha-2 (2024) two-letter country codes to represent countries, dependent territories, and special areas of geographical interest.

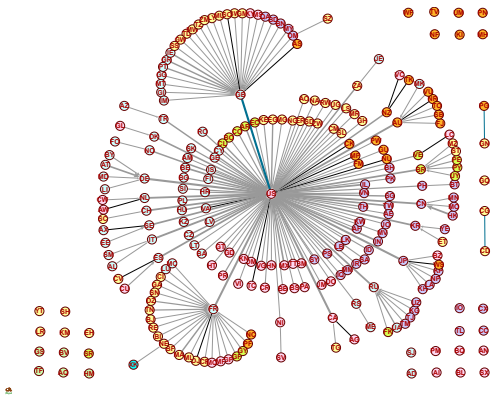
A problem in creating the co-authorship network between world countries is that the above query returns information about up to 200 most collaborative countries. The problem is resolved by considering the symmetry of the co-authorship data.

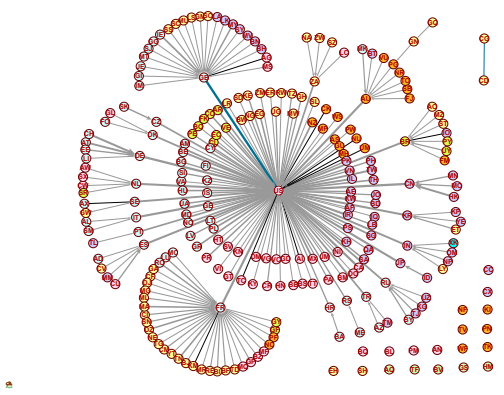
OpenAlex

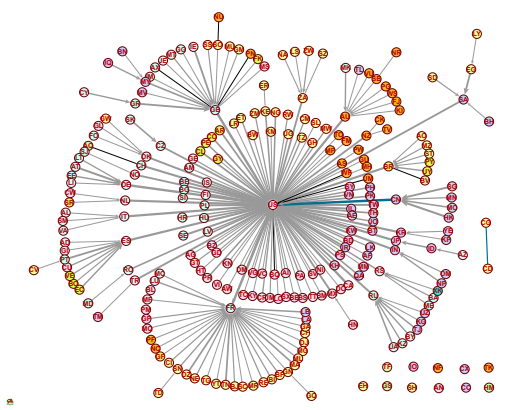
First neighbors

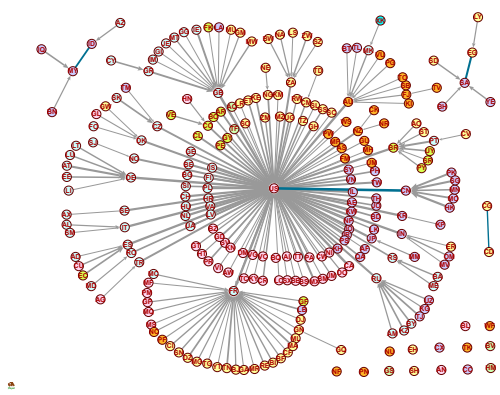


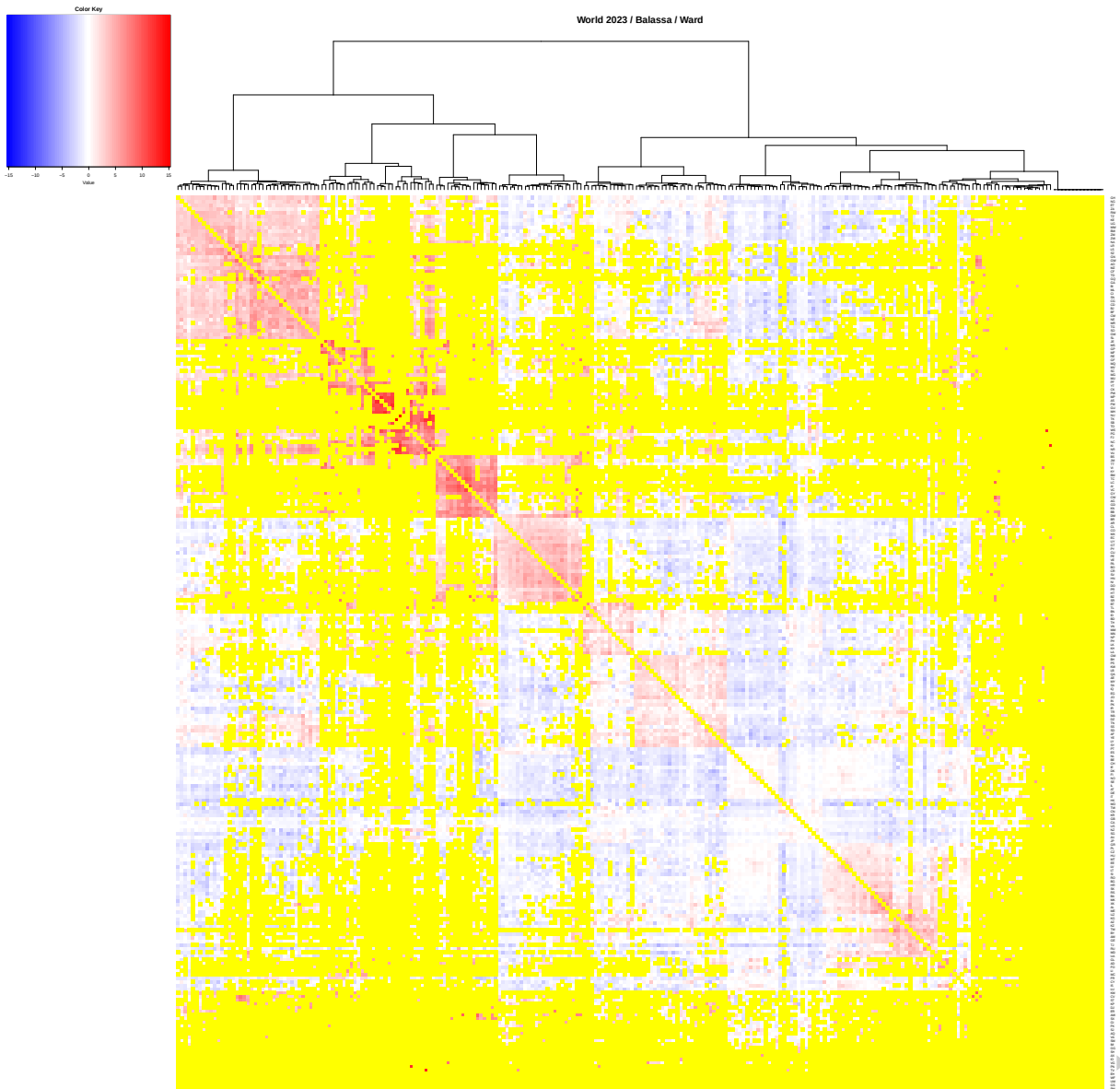












Conclusions

1. Open-houses
2. Webinars
3. Google user group
4. GitHub/topic/OpenAlex
5. [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).