OpenAlex2Pajek

The saturation approach to bibliographic networks construction

Vladimir Batagelj
IMFM, UP IAM

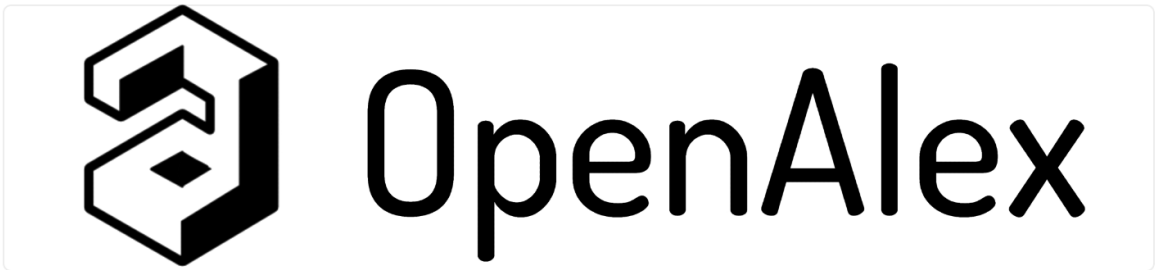
1345. + 1346. sredin seminar
Ljubljana, March 20 and March 27, 2024

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (March 27, 2024 at 17:07): 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

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).

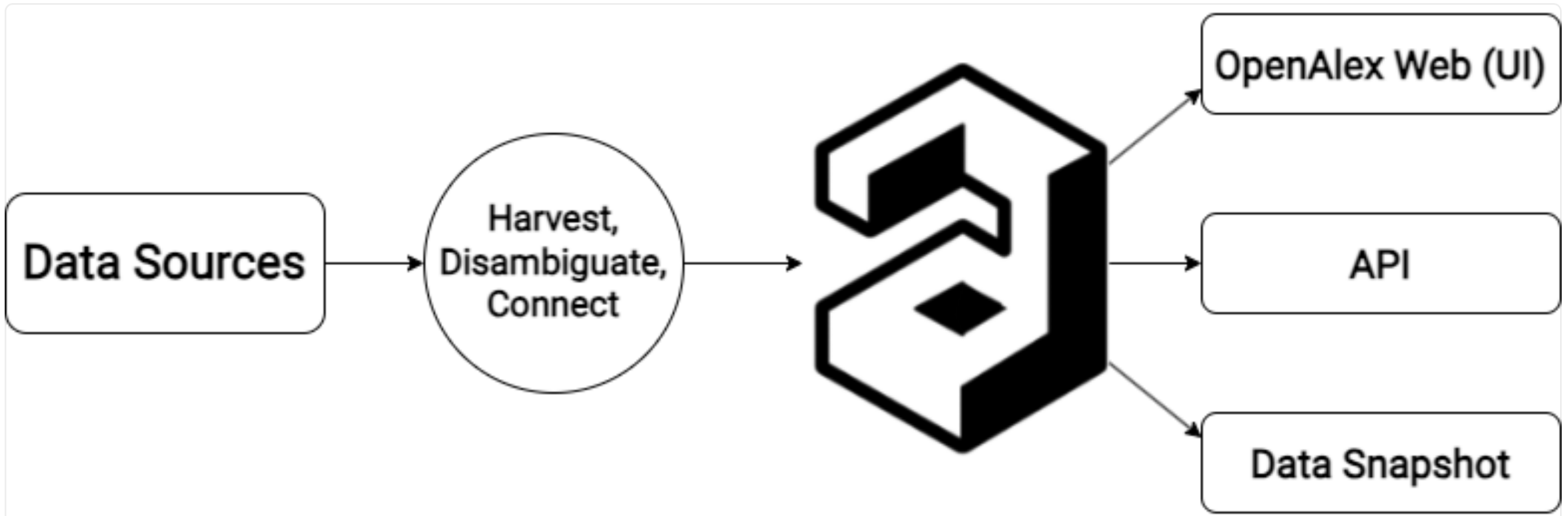
The problem of author countries – my example of an "extinct" country. W2033820728, W2059649701
(JSONview) Click API, see institutions

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

To ensure the OpenAlex longevity – UNESCO?

OpenAlex

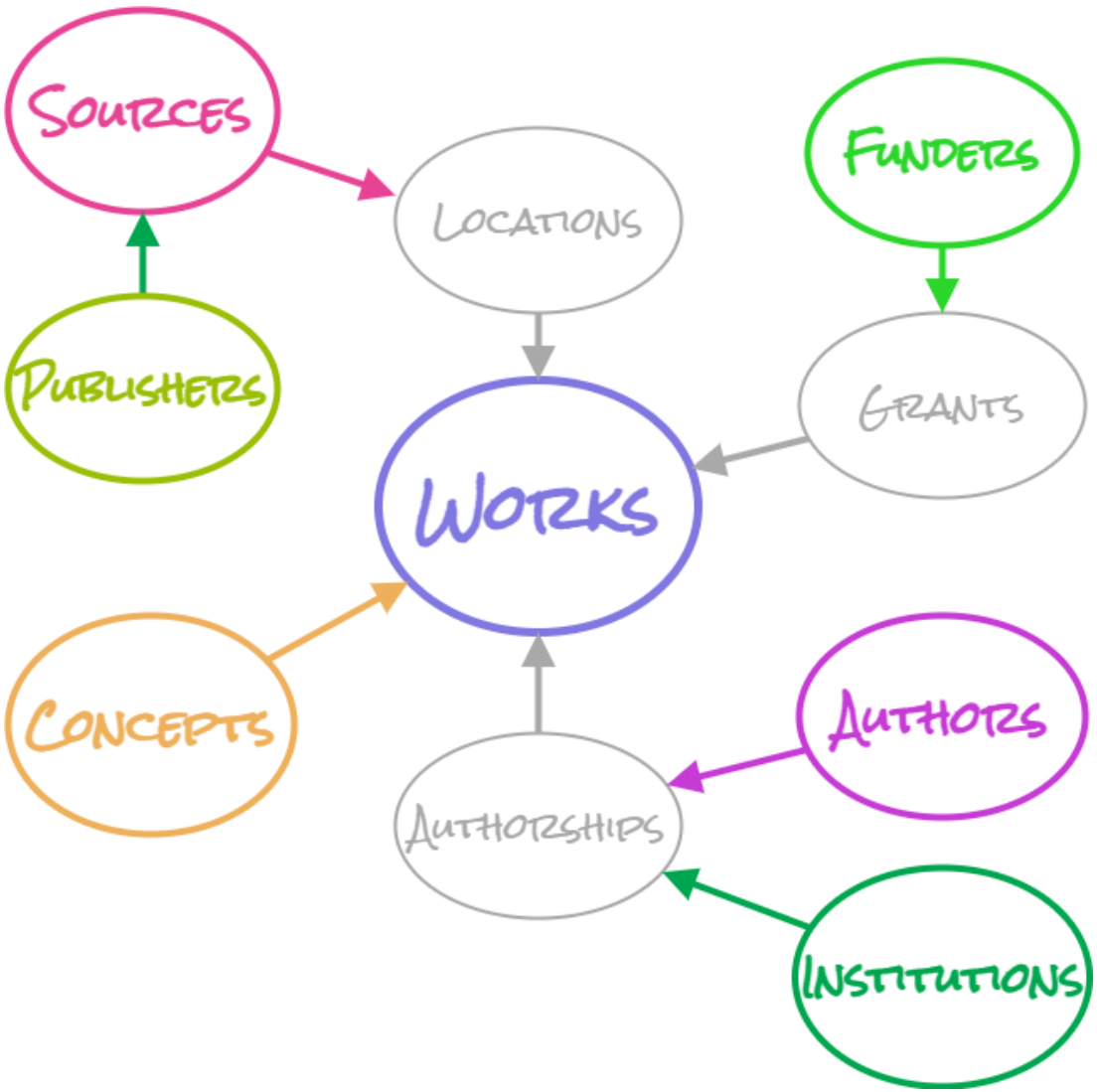
How it works



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

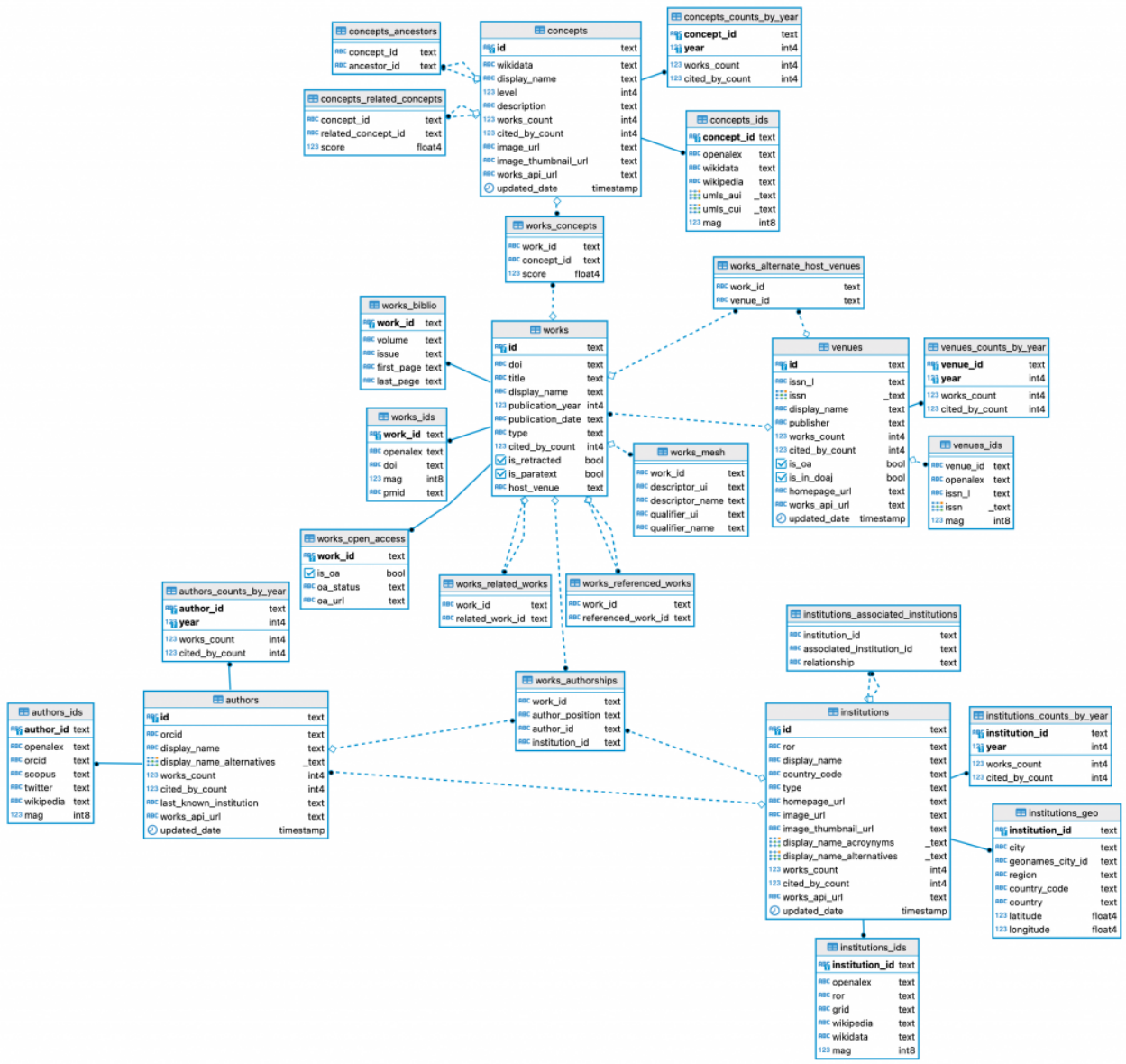
OpenAlex

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

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.

OpenAlex2Pajek

Saturation approach [3]

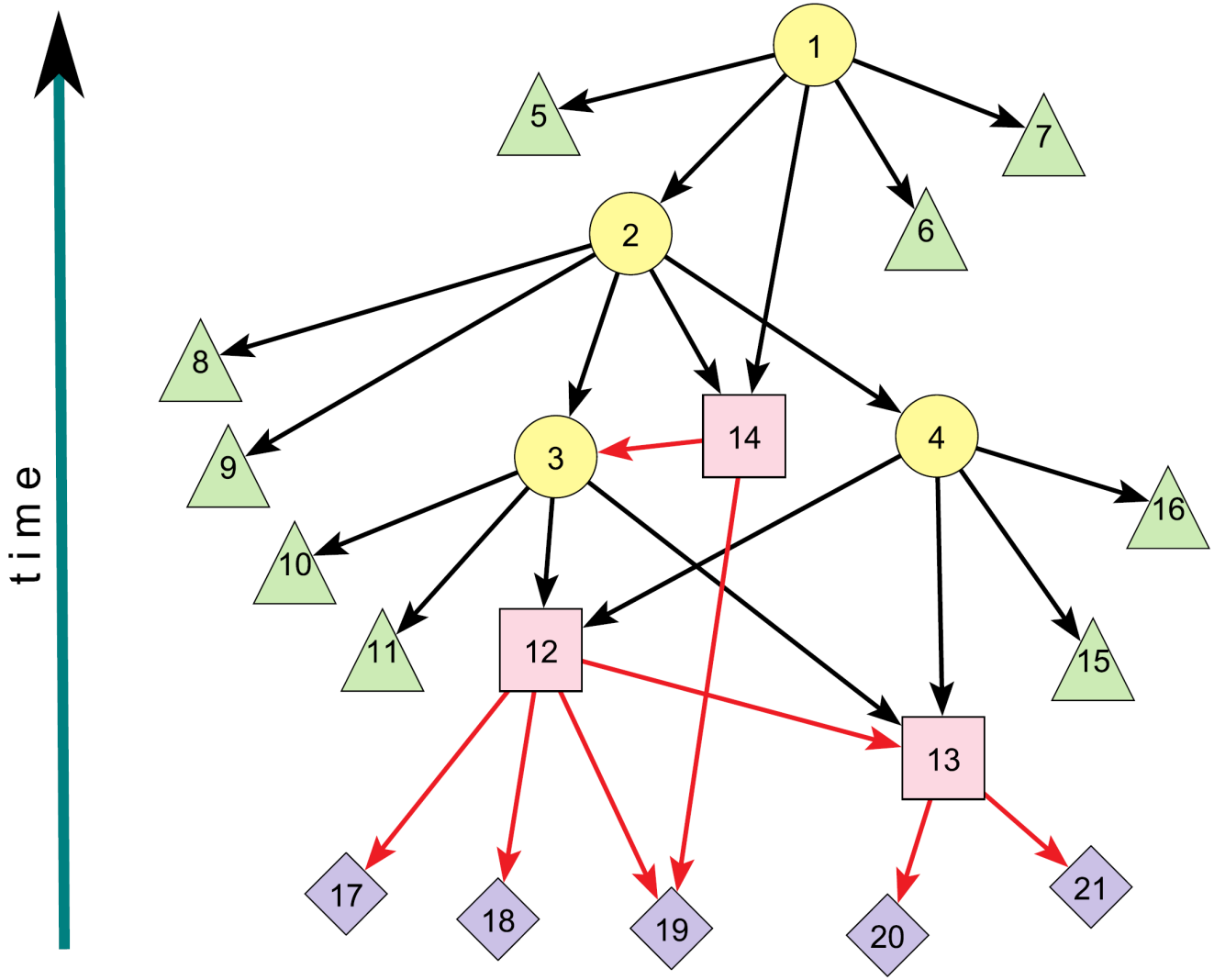
A first preliminary analysis performed in 2015 revealed that many works without a WoS description had large indegrees in the citation network. We manually searched for each of them (with indegree larger or equal to 20) and, when possible, we added them into the data set. It is important to note that earlier papers, which had a significant influence in the literature, did not often use the now established terminology (e.g., keywords) and were therefore overlooked by our queries.

After some iterations, we finally constructed the data set used in this paper. The final run of the program **`WoS2Pajek`** produced networks with sets of the following sizes: works

Considering the indegree distribution in the citation network **CiteAll**, we found that most works were referenced only once. Therefore, we decided to remove all 'only cited' nodes with indegree smaller than 3 ($DC = 0$ and $indeg < 3$)—the *boundary problem* (Batagelj et al. 2014). We also removed all only cited nodes starting with strings `''[ANONYM''`,

OpenAlex2Pajek

Saturation approach [3]



OpenAlex2Pajek

Saturation approach [3]

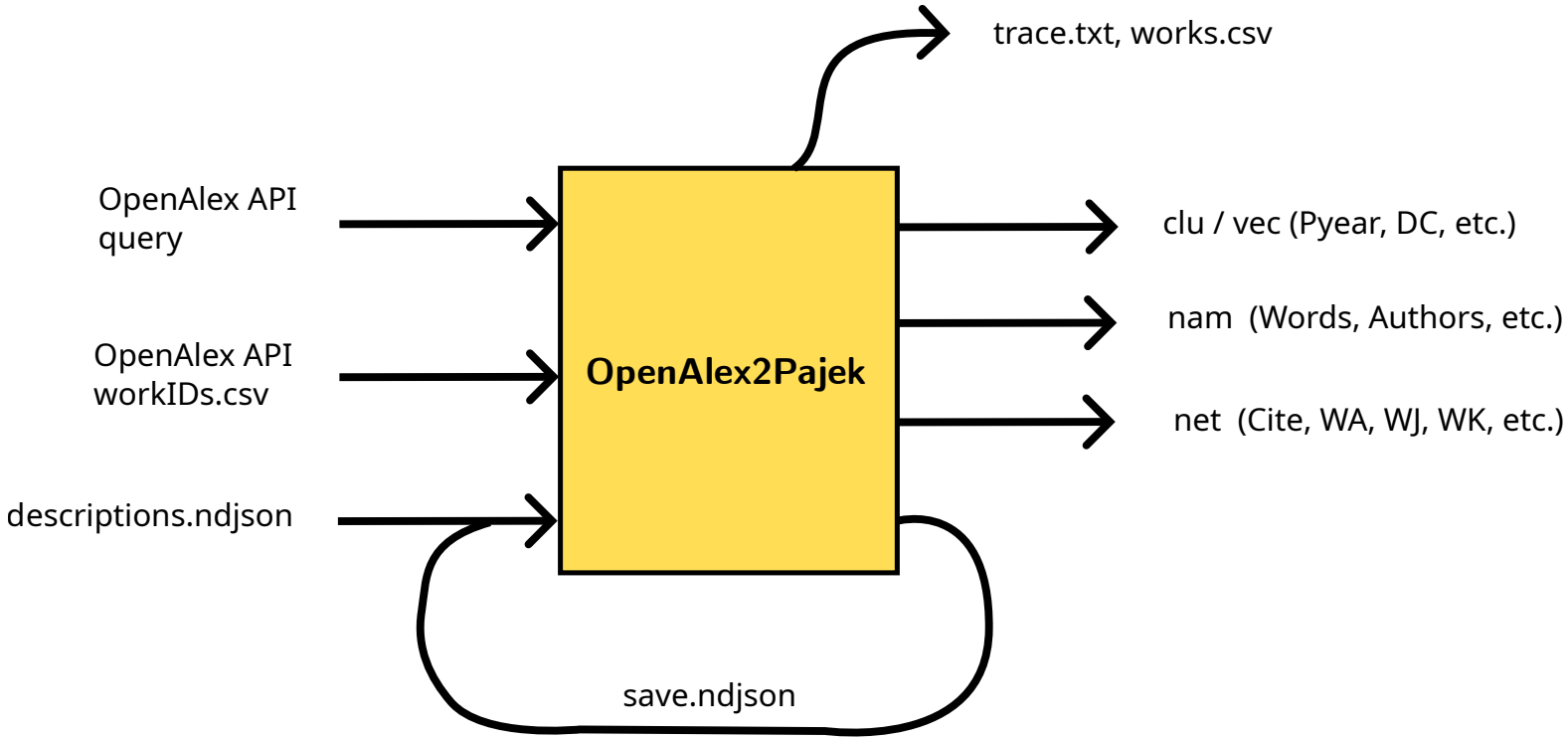
Figure 2 shows a schematic structure of a citation network. The circular nodes correspond to the query hits. The works cited in hits are presented with the triangular nodes. Some of them are in the following phase (search for often cited works) converted into the squares (found in WoS by our secondary search). They introduce new cited nodes represented as diamonds. It is important to note that the age of a work was determined by its publication year. In a citation network, in order to get a cycle, an "older" node had to cite a "younger or the same age" work. Given that this rarely happens, citation networks are usually (almost) acyclic.

The saturation procedure can be extended also in the other direction – to citing works.

W2083084326

OpenAlex2Pajek

Scheme / version 1



Processing a sequence of works

OpenAlex2Pajek

Main loop


```
source("OpenAlex.R")
save <- TRUE; step <- 500
if(save) json <- file("save.ndjson","w",encoding="UTF-8")
Q <- list( search="handball",
  # filter="publication_year:2015",
  select="id,primary_location,publication_year,publication_date,type,lan
  per_page="200"
)
openWorks(query=Q,list=NULL,file=NULL)
# openWorks(query=NULL,list=hiCitelist,file="saved.ndjson")
# openWorks(query=Q,list=hiCitelist,file=NULL)
cat("*** OpenAlex2Pajek - Start",date(),"\n"); flush.console()
repeat{
  w <- nextWork()
  if(is.null(w)) break
  if(save) write(toJSON(w),file=json)
  if(WC$n %% step==0) cat(date()," n =",WC$n,"\n"); flush.console()
  tryCatch(
    processWork(w),
    error=function(e){ cat("W",WC$n,w$id,"\n"); flush.console();
      print(e)} )
}
createNetworks()
closeWorks()
cat("*** OpenAlex2Pajek - Stop",date(),"\n"); flush.console()
```

OpenAlex2Pajek

Dictionaries in R / vectors

```
> vDict <- c(a=3,b="Ljubljana",c=3.14,d=FALSE,e=NA)
> vDict
          a           b           c           d           e
        "3" "Ljubljana"      "3.14"     "FALSE"          NA
> vDict["c"]
     c
"3.14"
> vDict["c"] <- 3.14529
> vDict["f"]
<NA>
  NA
> "a" %in% names(vDict)
[1] TRUE
> "z" %in% names(vDict)
[1] FALSE
> vDict["f"] <- 14
> vDict
          a           b           c           d           e           f
        "3" "Ljubljana"   "3.14529"     "FALSE"          NA        "14"
> length(vDict)
[1] 6
> vDict[!(names(vDict) %in% c('e'))]
          a           b           c           d           f
        "3" "Ljubljana"   "3.14529"     "FALSE"        "14"
```

OpenAlex2Pajek

Dictionaries in R / lists


```
> (lDict <- list(a=3,b="Ljubljana",c=3.14,d=FALSE,e=NA))
$a
[1] 3
$b
[1] "Ljubljana"
$c
[1] 3.14
$d
[1] FALSE
$e
[1] NA
> lDict[["c"]]
[1] 3.14
> lDict$b
[1] "Ljubljana"
> lDict[["f"]] <- c(5,9)
> lDict$f
[1] 5 9
> length(lDict)
[1] 6
> names(vDict)
[1] "a" "b" "c" "d" "e" "f"
> lDict[["e"]] <- NULL
> lDict
$a
[1] 3
$b
[1] "Ljubljana"
$c
[1] 3.14
$d
[1] FALSE
$f
[1] 5 9
```

OpenAlex2Pajek

Surviving function call

<<-

Environments:

```
eDict <- function() new.env(hash=TRUE,parent=emptyenv())
en <- eDict()
assign("key",value,env=en)
en[["key"]] <- value
exists("key",env=en,inherits=FALSE)
ls(en)
if(!is.null(v<-get0("key",envir=en))) {
  ## ... deal with v ...
}
rm("key",envir=en)
```

[4, 5]

[6]

```
keys = ls

eDict <- function(size=10000L) new.env(hash=TRUE,parent=emptyenv(),size=size)

getVals <- Vectorize(get,vectorize.args="x")

dict2DF <- function(dict,ind) {
  V <- as.data.frame(t(as.data.frame(getVals(keys(dict),dict))))
  for(n in colnames(V)) V[[n]] <- unname(unlist(V[[n]]))
  return(V[order(V[[ind]]),])
}

# Wid -> (wind,hit,cnt,inp,out)
putWork <- function(Wid,hit){
  if(exists(Wid,env=works,inherits=FALSE)){
    if(hit){ cat("W duplicate ",Wid,"\n",file=WC$tr)
      works[[Wid]][["cnt"]] <- works[[Wid]][["cnt"]]+1
    } else works[[Wid]][["inp"]] <- works[[Wid]][["inp"]]+1
  } else {
    works[[Wid]] <- list(wind=length(works)+1,cnt=0,inp=0,out=0)
    if(hit) works[[Wid]][["cnt"]] <- 1 else works[[Wid]][["inp"]] <- 1
  }
  return(works[[Wid]][["wind"]])
}

# Sid -> (sind)
putSrc <- function( ...

# Aid -> (aind)
putAuth <- function( ...
```

OpenAlex2Pajek

Program in R

First version: Approach from **WoS2Pajek**.
OpenAlex2Pajek
OpenAlex.R
OpenAlex2Pajek.R

Second version: Each OpenAlex bibliographic unit has a description (is a hit) $\Leftrightarrow$ in the first part determine the set of works W using the saturation approach and in the second part create **Pajek** networks (still in development).
OpenAlex2.R
CiteNet.R

OpenAlex2Pajek

Creating Handball Pajek networks

```
> source("CiteNet.R")
OpenAlex2Pajek - Start Wed Mar 27 11:43:34 2024
*** OpenAlex2Pajek - Start Wed Mar 27 11:43:34 2024
Wed Mar 27 11:43:42 2024  n = 500
Wed Mar 27 11:43:48 2024  n = 1000
Wed Mar 27 11:43:55 2024  n = 1500
...
Wed Mar 27 12:02:03 2024  n = 25000
Wed Mar 27 12:02:43 2024  n = 25500
*** OpenAlex2Pajek - Data Collected Wed Mar 27 12:03:05 2024
hits: 25905 works: 235710 authors: 52713 anon: 1331 sources: 5516
*** OpenAlex2Pajek - Stop Wed Mar 27 12:03:32 2024
> source("CiteNet.R")
OpenAlex2Pajek - Start Wed Mar 27 12:25:14 2024
*** OpenAlex2Pajek - Start Wed Mar 27 12:25:14 2024
Wed Mar 27 12:25:21 2024  n = 500
Wed Mar 27 12:25:27 2024  n = 1000
Wed Mar 27 12:25:35 2024  n = 1500
...
Wed Mar 27 12:44:43 2024  n = 26000
Wed Mar 27 12:47:31 2024  n = 26500
*** OpenAlex2Pajek - Data Collected Wed Mar 27 12:49:58 2024
hits: 26905 works: 244608 authors: 53882 anon: 1334 sources: 5558
*** OpenAlex2Pajek - Stop Wed Mar 27 12:50:27 2024
```

ZIP

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 [7]
7. [8], [9]

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. Batagelj, V., Ferligoj, A. & Squazzoni, F. The emergence of a field: a network analysis of research on peer review. *Scientometrics* **113,** 503–532 (2017).
4. Asai, S. *R6 Based Key-Value Dictionary Implementation.* https://cran.r-project.org/web/packages/Dict/index.html. https://github.com/five-dots/Dict. 2020.
5. Brown, C. & Hughes, J. *hash: Full Featured Implementation of Hash Tables/Associative Arrays/Dictionaries.* https://cran.r-project.org/web/packages/hash/. 2023.
6. Learning Machines. *Hash Me If You Can.* https://blog.ephorie.de/hash-me-if-you-can. 2019.

References II

7. Delgado-Quirós, L. & Ortega, J. L. Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 1–36 (2024).

8. Zhang, L., Cao, Z., Shang, Y., Sivertsen, G. & Huang, Y. Missing institutions in OpenAlex: possible reasons, implications, and solutions. *Scientometrics*, 1–23 (2024).

9. 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).