openalexR: An R-Tool for Collecting Bibliometric Data from OpenAlex

by Massimo Aria, Trang Le, Corrado Cuccurullo, Alessandra Belfiore, and June Choe

Abstract Bibliographic databases are indispensable sources of information on published literature. OpenAlex is an open-source collection of academic metadata that enable comprehensive bibliographic analyses (Priem, Piwowar, and Orr 2022). In this paper, we provide details on the implementation of openalexR, an R package to interface with the OpenAlex API. We present a general overview of its main functions and several detailed examples of its use. Following best API package practices, openalexR offers an intuitive interface for collecting information on different entities, including works, authors, institutions, sources, and concepts. openalexR exposes to the user different API parameters including filtering, searching, sorting, and grouping. This new open-source package is well-documented and available on CRAN.

1 Introduction

Bibliographic data sources are organized digital collections of reference metadata and citation links related to published scientific literature. One primary purpose is to assess the scholarly performance, although it is important to exercise caution when employing bibliometric indicators for performance evaluation (Van Noorden, 2013; Hicks et al., 2015; Priem et al., 2011). Another objective of bibliometric analyzes is science mapping, a process that involves synthesizing extensive data, prioritizing impactful research, and extracting key knowledge structures (Chen, 2017). Given the rapid proliferation of scientific publications, science mapping evidence is particularly valuable for reconstructing the theoretical framework of empirical studies or literature reviews.

Recently, new sources of multidisciplinary bibliographic data have emerged, including Microsoft Academic, launched in 2016 (Sinha et al., 2015; Wang et al., 2019, 2020), Semantic Scholar launched in 2015 (Ammar et al., 2018), and CrossRef, an open bibliographic data source launched in 2017 (Hendricks et al., 2020; Van Eck et al., 2018). Dimensions is a scientometric data source that provides also information on grants, datasets, clinical trials, patents, and policy documents (Herzog et al., 2020; Hook et al., 2018). These databases complement the many existing open and commercial sources. The two most widely used commercial multidisciplinary databases are Web of Science and Scopus while the specialized ones include PubMed, EconBiz, and arXiv, which are the main open bibliographic sources for medicine, economics, and physical sciences and engineering, respectively.

The value of these open and commercial bibliographic data sources hinges on several characteristics: (1) document coverage, (2) completeness and accuracy of citation links, (3) update speed, (4) automation of data access through web interfaces, APIs, and data dumps, and (5) the terms of use for a data source (Wanyama et al., 2022; Kulkanjanapiban and Silwattananusarn, 2022; Singh et al., 2021; Martín-Martín et al., 2021; Visser et al., 2021; Waltman and Larivière, 2020; Winter, 2017). OpenAlex is recognized for providing the most extensive coverage of scientific literature, encompassing a notably larger number of documents compared to other data sources (see https://openalex.org/about). Its document coverage outpaces all major databases, including Microsoft Academic (Visser et al., 2021; Wang et al., 2020) and CrossRef, which are its primary sources. While Google Scholar reportedly boasts an estimated 389 million database, it is crucial to note that Google Scholar does not adhere to the conventional database model due to its lack of comprehensive metadata, and it does not permit users to download query search results (Dallas et al., 2018).

Another remarkable strength of OpenAlex is its substantial collection of open-access works, totaling 48 million. This extensive repository grants users open and free access to a wealth of scholarly resources, aligning with OpenAlex's dedication to open science principles. This commitment promotes both accessibility and transparency in the sharing of knowledge. Furthermore, OpenAlex stands out not only in terms of quantity but also in data quality. With a substantial number of citations amounting to 1.9 billion, OpenAlex emphasizes its relevance for researchers engaged in citation-based studies. This robust citation data enhances its appeal as a valuable resource for scholars conducting research reliant on citation analysis.

The purpose of this article is to introduce the OpenAlexR R package, which facilitates the retrieval of metadata from OpenAlex, performs specific functions, and formats the data for utilization in bibliometrix, particularly for science mapping and research assessment purposes.

1.1 OpenAlex

Named after the ancient Library of Alexandria, OpenAlex is a free and fully open catalogue of scholarly metadata with open data, open APIs, open-source code (Priem et al., 2022). Behind this tour de force is OurResearch, a nonprofit organisation dedicated to open principles of academic works with other impactful projects such as CiteAs (Du et al., 2021) and Unpaywall (Chawla, 2017). OpenAlex was launched in January 2022, timely replacing the retired Microsoft Academic system. OpenAlex is already extensively used in many scholarly articles (Belfiore et al., 2022). The data is currently accessible through a complete database snapshot and a REST API that is updated daily.

The OpenAlex database consists of eight academic entity types: works, authors, institutions, sources, concepts, publishers, funders, geo (Fig. 1). It is important to know the OpenAlex entities because it is possible to make query for each entity. In fact, each entity is assigned an OpenAlex ID (OAID) which represents the primary key to access the data. However, OpenAlex also recognizes different external canonical IDs for different entities. We briefly summarise the eight entities below. For more detail, visit the documentation page and the more recent Postgres schema diagram by OpenAlex.

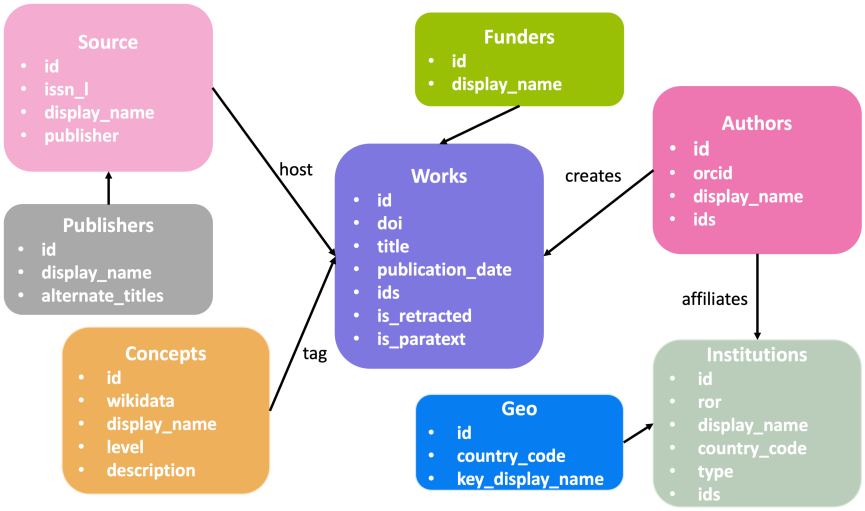

Figure 1: Eight OpenAlex entities and their first few attributes.

- **Works**: academic documents such as journal articles, proceedings, books, and datasets. The Works entity is central in that it ties together the other four entities (Fig. 1).
  OpenAlex indexes over 200 million works. One can identify a work by its OAID or DOI, a unique alphanumeric string assigned to a digital document, often a research article.
- **Authors**: individuals who create works. Authors create Works, study Concepts, and are affiliated with Institutions. OpenAlex indexes more than 200 million authors. One can identify an author by their OAID or ORCID, a persistent and unique identifier assigned to researchers.
- **Institutions**: universities and organisations affiliated with authors. Institutions are linked to Works via Authors. OpenAlex indexes more than 100,000 institutions. One can identify an institution by its OAID or ROR ID, a persistent identifier for research organizations.
- **sources**: repositories that house works such as journals, conferences, preprint repositories, or institutional repositories. OpenAlex uses a fingerprinting algorithm to match multiple locations a work may be hosted in and flag the version of the record's host as primary. OpenAlex indexes more than 100,000 sources. One can identify a source by its OAID or ISSN-L, *i.e.*, a single ISSN that groups the publication's all possible ISSNs (standardized numeric identifiers assigned to serial publications).
- **Concepts**: topics of works, deduced from their titles and abstracts. To identify a concept, you can use the OAID or its Wikidata ID, a unique identifier assigned to that entity in Wikidata because all OpenAlex concepts are also Wikidata concepts. The concepts follow a hierarchy; there are 19 concepts at the root level (0-level) and 5 layers of descendants. OpenAlex indexes $\sim$ 65,000 concepts.
- **Publishers**: companies and organizations that distribute academic documents. Each publisher publishes multiple journals, so publisher data is aggregate data. OpenAlex indexes about 10,000 publishers.

- **Funders**: public and private companies that fund research. OpenAlex indexes about 30,000 funders. Funder data comes from Crossref, and is enhanced with data from Wikidata and ROR.
- **Geo**: works produced in the country based on the nationality of the institution with which the author is affiliated. In particular, there are some ways to filter and group academic documents by continents and the Global South. OpenAlex uses United Nations data to divide the globe into continents and regions, making it easier to filter the data.

2 Implementation of openalexR

Interacting with the OpenAlex API, **openalexR** provides easy querying and downloading of scholarly metadata as well as converting the output into a classic bibliographic dataframe (Fig. 2), which allows the user to streamline their data collection and downstream analyses (Aria, 2022).

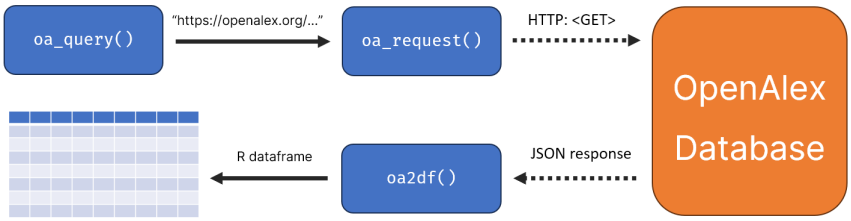

Figure 2: openalexR workflow

With minimal dependencies, **openalexR** lowers the barrier to using the REST API by simplifying input entry, handling rate limits, and automatically parsing API responses. The package also offers other useful functionality such as snowball search and N-grams of works. **openalexR** is currently listed on the OpenAlex website as the supported R library for API access. The main function of **openalexR**, `oa_fetch`, is a convenience wrapper for three smaller functions to

1. generate a valid query from a set of arguments provided by the user (`oa_query`),
2. download a collection of entities matching the query (`oa_request`),
3. and convert the list output to a classical bibliographic data frame (`oa2df`).

Specifically, in constructing valid queries following the OpenAlex API syntax, `oa_query` utilizes the `modify_url` function of the **httr** package (Wickham, 2022) to pass the `filter`, `sort`, `search`, and `group_by` parameters specified by the user. Next, `oa_request` sends a request to OpenAlex, downloads the JSON output matching the created query, and returns the result in a nested list. Finally, `oa2df` converts this output list into a classic bibliographic data frame (similar to an Excel sheet) that can be used as input in a bibliometric analysis or scientific mapping (Wais, 2016), *e.g.*, using the **bibliometrix** package (Aria and Cuccurullo, 2017).

In addition to `oa_fetch`, the package **openalexR** includes three other functions specific to certain features: `oa_random`, `oa_snowball`, and `oa_ngrams`. The function `oa_random`, similar to `oa_fetch`, returns a record randomly. This function is particularly useful for casual sampling purposes. For instance, when conducting an analysis of gender bias within the academy, one could utilize this function to randomly query the database.

`oa_snowball` enables the user to perform *snowballing*. Snowballing, or snowball search, is a literature search technique where the researcher starts with a set of articles and finds other articles that cite (forward citations) or were cited by (backward citations) the original set (Wohlin, 2014). Meta-analysis researchers often employ this technique to collect relevant primary studies, where sufficient iterations of snowballing can converge on and exhaust the target literature space (Siddaway et al., 2019). The traditional method of conducting a snowball search involves manual effort. However, computer-assisted snowball search can effectively reduce the time and resources required while maintaining a representative coverage of the target literature (McWeeny et al., 2021, 2022). `oa_snowball` takes a vector of OpenAlex IDs as input and returns a list of 2 elements: `nodes` and `edges`. `oa_snowball` locates and retrieves information on articles that cite or are cited by the initial set of articles (`nodes`) and also the relationships between these articles (`edges`). Following **tidygraph**'s convention (Pedersen, 2022b), the `edges` dataframe contains 2 columns of OpenAlex IDs, *from* and *to*. The row *from A to B* means *A cites B*.

The `oa_ngrams` function is used to obtain N-grams from a specific set of works. N-grams are defined as sequences of n words that occur within a work and are commonly used in language

modeling and text analysis to identify relationships between words. The use of N-grams allows for the analysis of the frequency and distribution of words within a text or set of texts and is widely employed in fields such as natural language processing, machine translation, and sentiment analysis. Some work entities in OpenAlex include N-grams of their full text, which are obtained from the Internet Archive using the spaCy parser to index academic works. The **openalexR** package offers the capability to extract N-grams of works through the `oa_ngrams` function. This function takes a vector of OpenAlex IDs as input and returns a list of N-grams and their corresponding frequencies.

3 Installation of openalexR

The package is available from the Comprehensive R Archive Network (CRAN) via the command `install.packages("openalexR")`. The current version is v1.2.0. Development versions (latest v1.2.0.9000) are available on GitHub and can be installed using **devtools** (Wickham et al., 2022) or **remotes** (Csárdi et al., 2021).

```
install.packages("devtools")
devtools::install_github("ropensci/openalexR")
or
install.packages("remotes")
remotes::install_github("ropensci/openalexR")
```

Other installation details are available on the GitHub page https://github.com/ropensci/openalexR.

4 Polite use

The OpenAlex API doesn't require authentication but requires following polite usage. To get into the polite pool, it is necessary to provide a user e-mail address through the `mailto` parameter in R options

```
options(openalexR.mailto = "example@email.com")
```

in all API requests. The polite pool has much faster and more consistent response times.

5 Examples of use

We show many different examples of typical use cases on the package's README and vignettes. Examples we show in this manuscript can be found at https://github.com/trangdata/oarj/blob/main/paper-examples.md.

First, we demonstrate an example that uses a few different filters. We want to describe the use of "*bibliometrics*" approaches in the scientific literature. We first describe how to make the query that allows us to answer this search question. After a brief description of the concept "*bibliometrics*", we identify the scientific literature that has used the concept "*bibliometrics*" in OpenAlex. Finally, focusing on the metadata offered by OpenAlex, we analyse the most relevant sources, authors, institutions, and works on bibliometrics.

5.1 The *bibliometrics* concept

We define the search on the entity "concepts" by filtering the "bibliometrics" topic associated with the id C178315738. The function `oa_fetch` generates the query from the set of arguments provided to it, downloads the set of concepts that match the query, and converts the output into a classical bibliographic data frame. Concepts can be queried using concept IDs or by concept name searching.

Searching "bibliometrics" concept by name:

```
concept <- oa_fetch(
  entity = "concepts",
  display_name.search = "bibliometrics" # search by concept name "bibliometrics"
)

concept$id
# [1] "https://openalex.org/C178315738"
```

```
cat(concept$description, "\nis a level", concept$level, "concept")
# [1] statistical analysis of written publications, such as books or articles
is a level 2 concept
```

Alternatively, once we know the OAID for a concept, we can search by its ID and get a similar result:

```
concept <- oa_fetch(
  entity = "concepts",
  identifier = "C178315738" # OAID for "bibliometrics"
)

cat(concept$description, "\nis a level", concept$level, "concept")
# [1] statistical analysis of written publications, such as books or articles
is a level 2 concept
```

To describe which concepts are related to the term *bibliometrics*, let's analyze the OpenAlex hierarchy. In OpenAlex each work is tagged with multiple concepts, based on the title, abstract and host source title. A score is available for each concept in a work, demonstrating how well that concept represents the work to which it was assigned. However, when a lower-level descendant concept is assigned, all of its antecedent concepts are also assigned. Since *bibliometrics* is a level 2 concept in OpenAlex, we have detailed information on the concepts related to ancestors (level 0 or 1), peers (level 2), and descendants (level 3).

```
related_concepts <- concept$related_concepts[[1]] |>
  dplyr::mutate(relation = case_when(
    level < 2 ~ "ancestor",
    level == 2 ~ "equal level",
    TRUE ~ "descendant"
  )) |>
  dplyr::arrange(level) |>
  dplyr::relocate(relation) |>
  dplyr::select(-wikidata)

# output in Table 1:
related_concepts
```

We find 4 ancestor, 11 equal-level and 9 descendant concepts of *bibliometrics* (Tab. 1). The resulting hierarchy of *bibliometrics* can enable us to analyse, for example, all equal-level concepts.

```
concept_df <- oa_fetch(
  entity = "concepts",
  identifier = c(concept$id, equal_level$id)
)

concept_df |>
  dplyr::select(display_name, counts_by_year) |>
  tidyr::unnest(counts_by_year) |>
  dplyr::filter(year < 2022) |>
  ggplot(aes(x = year, y = works_count, color = display_name)) +
  facet_wrap(~display_name) +
  geom_line() +
  ...
```

Visualising all *bibliometrics*-related concepts together, we observe an increasing trend in the subfields of *bibliometrics*, *peer review*, *scientific literature* and *scientometrics*. Conversely, there has been a reduction in the number of papers in the subfields of *webometrics*, *knowledge organisation*, *information science*, *collection development*, and *altmetrics*. *PageRank* and *Impact factor*, concepts have remained stable in popularity over the last 10 years. Compared to other topics, *citation* has the highest number of papers over time (Fig. 3).

5.2 *Bibliometrics* dataset

We download all works, included in OpenAlex, that have the words *bibliometrics* or *science mapping* in the title to map bibliometric approaches in the scientific literature. We set the query using the

relation	id	display_name	level	score
ancestor	C124101348	Data mining	1	
ancestor	C161191863	Library science	1	
ancestor	C136764020	World Wide Web	1	
ancestor	C41008148	Computer science	0	
equal level	C525823164	Scientometrics	2	6.6193560
equal level	C2779455604	Impact factor	2	4.1035270
equal level	C2778407487	Altmetrics	2	2.5396087
equal level	C521491914	Webometrics	2	2.3026270
equal level	C2781083858	Scientific literature	2	1.6163236
equal level	C2778805511	Citation	2	1.6110690
equal level	C95831776	Information science	2	1.5750017
equal level	C2779172887	PageRank	2	1.5363927
equal level	C138368954	Peer review	2	1.4112837
equal level	C2779810430	Knowledge organization	2	1.0037539
equal level	C2780416505	Collection development	2	0.8137859
descendant	C105345328	Citation analysis	3	4.9036117
descendant	C2778793908	Citation impact	3	4.0405297
descendant	C2780378607	Informetrics	3	2.1396947
descendant	C2778032371	Citation index	3	1.8888942
descendant	C83867959	Scopus	3	1.6536747
descendant	C2776822937	Bibliographic coupling	3	1.3375385
descendant	C2779693592	Journal ranking	3	1.1321522
descendant	C45462083	Documentation science	3	0.8473609
descendant	C2777765086	Co-citation	3	0.8002241

Table 1: Concepts related to *bibliometrics*: ancestors, equal-levels, and descendants.

following parameters: `entity` is "works"; `title.search` is `"bibliometrics|science mapping"` and, for the first part, `count_only` is `TRUE` so we can see how many records will be returned.

```
oa_fetch(
  entity = "works",
  title.search = "bibliometrics|science mapping",
  count_only = TRUE
)
#      count db_response_time_ms page per_page
# [1,] 26,953               118    1        1

biblio_works <- oa_fetch(
  entity = "works",
  title.search = "bibliometrics|science mapping",
  count_only = FALSE
)
```

Our query returns 26,953 works concerning bibliometrics. By default, the `oa_fetch` converts these records in a tibble object (dataframe) with each row containing information about a work. This dataframe has 28 columns containing important information about a work, such as the publication date, DOI, reference works, and so on. If users wish to convert the original nested list into another object, they can change the parameters in the following way `oa_fetch(..., output = "list")`. From this dataset, we could describe the most relevant sources, authors, and institutions.

5.3 Most relevant sources

We first identify the core journals for the discipline by tallying all bibliometrics-related works for each source and selecting the top 5 sources with the most works.

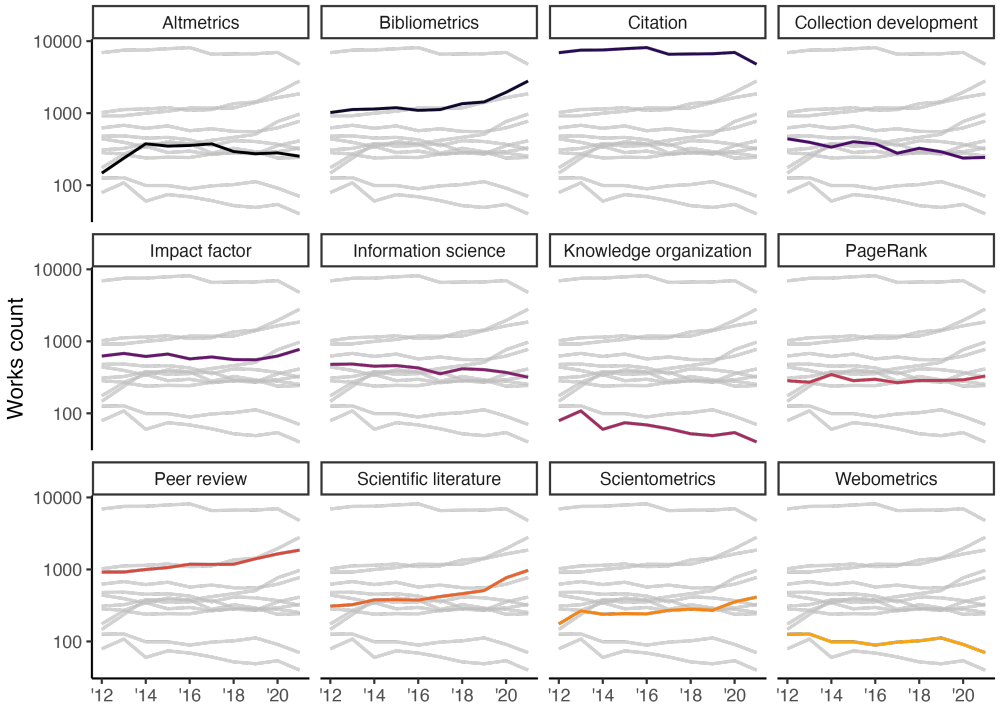

Figure 3: Trends in *bibliometrics*-related topics in the past 10 years.

```
sources <- biblio_works |>
  dplyr::count(so) |>
  tidyr::drop_na(so) |>
  dplyr::slice_max(n, n = 5) |>
  dplyr::pull(so)

sources
# [1] "Scientometrics"
# [2] "Sustainability"
# [3] "Social Science Research Network"
# [4] "International Journal Environmental Research and Public Health"
# [5] "Environmental Science and Pollution Research"
```

Visualising these counts over the years (Fig.4), we observe the field has expanded overall, especially starting around 2015. *Scientometrics* is the oldest journal publishing on bibliometrics and remains the top source for these articles. Other journals started to publish these works in 2015 and have maintained some volume in this field. *Sustainability* only started publishing in this field in 2017 but has rapidly increased its number of publications since. In 2021, it was the second source to have published the most bibliometrics articles, after *Scientometrics*.

5.4 Most relevant authors and institutions

Secondly, we identify the authors and institutions most relevant to the discipline by extracting the list of author and institution-related metadata from the collection, tallying all bibliometrics-related works for each author, and selecting the top 10 authors who write the most articles and 10 institutions with the most publications in the field.

```
biblio_authors_raw <- do.call(rbind.data.frame, biblio_works$author)
biblio_insts <- biblio_authors_raw |>
  dplyr::count(institution_display_name) |>
  dplyr::rename("name" = institution_display_name) |>
  tidyr::drop_na(name) |>
  dplyr::slice_max(n, n = 10) |>
  dplyr::mutate(type = "Institution")
```

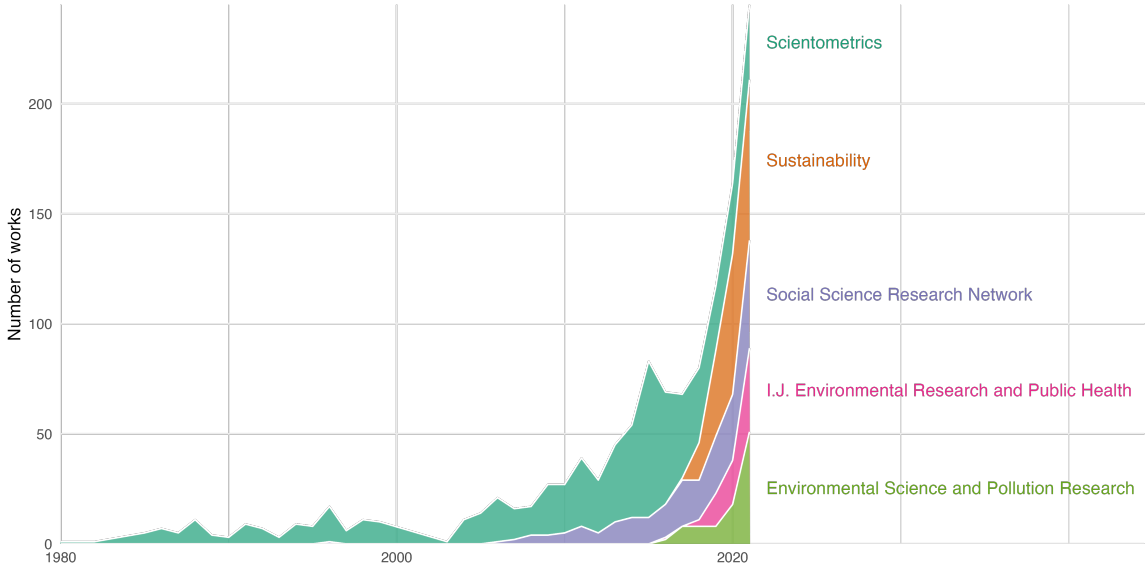

Figure 4: Number of *bibliometrics* articles by journal over the years.

```
biblio_authors <- biblio_authors_raw |>
  dplyr::count(au_display_name) |>
  dplyr::rename("name" = au_display_name) |>
  tidyr::drop_na(name) |>
  dplyr::slice_max(n, n = 10) |>
  dplyr::mutate(type = "Author")
```

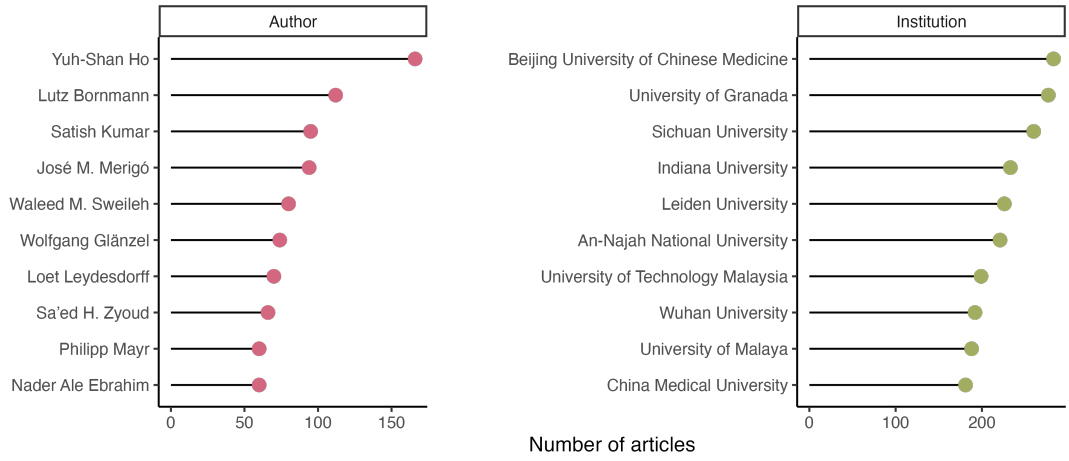

Figure 5: Most relevant authors and institutions.

Among the 84,335 authors in our collection, Yuh-Shan Ho appears to have published the most bibliometrics articles. He has over 150 bibliometric papers in his career. Another relevant author in our collection is Lutz Bornmann with more than 100 bibliometrics papers. The other authors in the top 10 all have between 50 and 100 papers (Fig. 5).

With regard to institutions, we observe that the concept *bibliometrics* is widely developed by different centres of expertise. At the top of the ranking, we see Beijing University of Chinese Medicine (China) and University of Granada (Spain) with over 250 articles. Other relevant institutions in our dataset include Sichuan University (China), Indiana University (US), Leiden University (Netherlands), An-Najah National University (Palestine) with more than 200 bibliometric papers. Other institutions in the top 10 have between 150 and 200 papers (Fig.5).

5.5 Most relevant works

We identify the most relevant papers for the discipline by tallying all citations of articles related to bibliometrics and selecting the top 10 most cited articles (Tab.2). We find 'Software survey: VOSviewer, a computer programme for bibliometric mapping' at the top with 4805 citations. Other relevant works in our collection are 'Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses' with 1996 citations and 'bibliometrix: An R-tool for comprehensive science mapping analysis' with 1800 citations. The other papers in the top 10 all have between 950 and 1500 citations.

```
seminal_works <- slice_max(biblio_works, cited_by_count, n = 10)

# output in Table 2:
seminal_works |>
  dplyr::select(publication_date, display_name, so, cited_by_count)
```

Year	Article	source	Cited
2010	Software survey: VOSviewer, a computer program for bibliometric mapping	Scientometrics	5364
2017	bibliometrix : An R-tool for comprehensive science mapping analysis	Journal of Informetrics	2127
1976	A general theory of bibliometric and other cumulative advantage processes	Journal of the American Society for Information Science	1505
2015	Bibliometric Methods in Management and Organization	Organizational Research Methods	1488
2015	Bibliometrics: The Leiden Manifesto for research metrics	Nature	1168
2011	Science mapping software tools: Review, analysis, and cooperative study among tools	Journal of the Association for Information Science and Technology	1100
2004	Changes in the intellectual structure of strategic management research: a bibliometric study of the Strategic Management Journal, 1980–2000	Strategic Management Journal	1038
2010	A unified approach to mapping and clustering of bibliometric networks	Journal of Informetrics	922
2015	Green supply chain management: A review and bibliometric analysis	International Journal of Production Economics	920
2006	Forecasting emerging technologies: Use of bibliometrics and patent analysis	Technological Forecasting and Social Change	804

Table 2: Most relevant works

5.6 Snowball search

We perform snowballing with `oa_snowball` to identify the set of articles that cite and are cited by the two seminal works associated with the concept of bibliometrics: Software survey: VOSviewer, a computer programme for bibliometric mapping (W2150220236), bibliometrix : An R-tool for comprehensive science mapping analysis (W2755950973). We insert these OAIDs as identifiers in `oa_snowball`, use the filter on the citations obtaining only those related to 2022, then use **tidygraph** (Pedersen, 2022b) and **ggraph** (Pedersen, 2022a) to display this citation network (Fig. 6). `oa_snowball` returns a list of 2 elements: nodes and edges. The first have information about the work, while edges have the start and end points of the links. This list output from `oa_snowball` can be used directly as input to standard graph functions such as `tidygraph::as_tbl_graph` for further network analyses and visualisations in co-citation analysis, historiograph analysis, *etc.*

```
sb_docs <- oa_snowball(
  identifier = c("W2150220236", "W2755950973"),
  citing_filter = list(from_publication_date = "2022-01-01")
)

# Reduced output
print(sb_docs)
$nodes
# A tibble: 5,769 × 37
  id          display_name
  <chr>       <chr>
1 W2150220236 Software survey: VOSviewer, a computer program for bibliometric ...
2 W2755950973 bibliometrix : An R-tool for comprehensive science mapping analysis
3 W4306178549 Literature reviews as independent studies: guidelines for academic ...
```

```
4 W4320070415 Is Metaverse in education a blessing or a curse: a combined content  ...
# i 5,765 more rows
# i 35 more variables
$edges
# A tibble: 6,444 × 2
  from        to
  <chr>       <chr>
1 W4306178549 W2755950973
2 W4320070415 W2150220236
3 W3203542139 W2150220236
4 W4283392904 W2150220236
# i 6,440 more rows

# Conversion to a `tbl_graph` object for network analysis and visualization
sb_docs_graph <- tidygraph::as_tbl_graph(sb_docs)
```

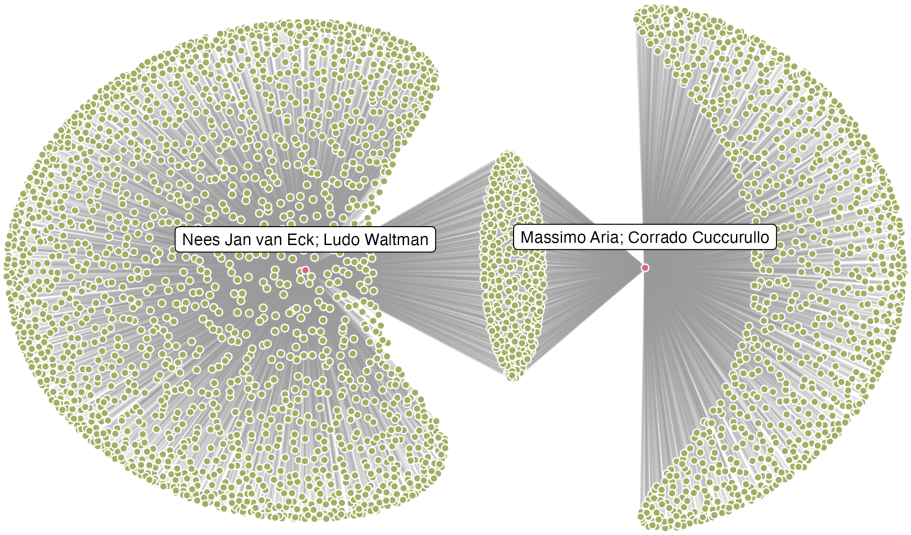

Figure 6: Two seminal works and their citations and references, from `oa_snowball` output.

`oa_snowball` finds and returns the metadata on the two seminal works in our dataset, and also information on the articles that cite and are cited by them, in 2022. We have a total of 2,907 citing articles: 2,078 articles cite van Eck et al., 1,161 articles cite Aria and Cuccurullo. In addition, as we can see from the graph, there are 332 articles citing both van Eck et al. and Aria and Cuccurullo, simultaneously.

5.7 N-grams

Finally, we obtain the N-grams of all the bibliometric works that make up our collection. N-grams are groups of words that occur in the full text of a work. To extract n-grams we use the `oa_ngrams` function from the **openalexR** package. From this list we then extract only the bigrams because we believe they can be more informative.

```
ngrams_data <- oa_ngrams(sample(biblio_works$id, 1000), verbose = TRUE)
top_10 <- do.call(rbind.data.frame, ngrams_data$ngrams) |>
  dplyr::filter(ngram_tokens == 2, nchar(ngram) > 10) |>
  dplyr::arrange(desc(ngram_count)) |>
  dplyr::slice_max(ngram_count, n = 10, with_ties = FALSE)
```

As can be seen from the graph of the 10 most frequent bi-grams, the papers are very much focused on the use of advanced technologies to improve efficiency and sustainability in various fields, such

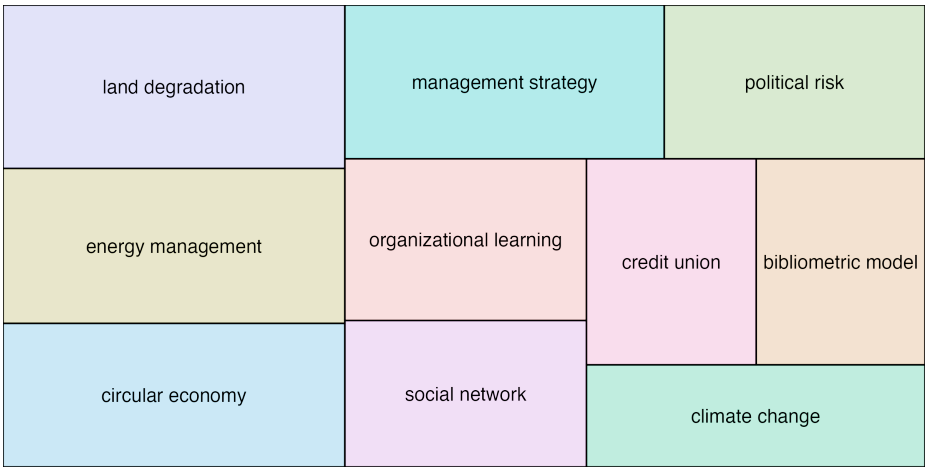

Figure 7: TreeMap of the top 10 bi-grams of *bibliometrics* articles.

as *circular economy*, *climate change*, *management strategy*, and *energy management* (Fig. 7). In addition, there are themes related to *political risk*, *organizational learning*, and cooperation within organizations and *social networks*. In general, these topics suggest the need for greater attention to environmental and social issues in business management and risk prevention. Finally, *bibliometric model* (scientific performance evaluation) and *credit union* (financial regulation) are interrelated to bibliometrics, which requires a global vision and cooperation among different stakeholders to be effectively applied.

6 Summary

openalexR is a new package that facilitates querying, collecting, and downloading bibliographic metadata of OpenAlex entities through the provided REST APIs. It is available on CRAN at https://cran.r-project.org/package=openalexR. **openalexR** helps streamline the researcher's workflow in accessing, collecting, and wrangling OpenAlex data. Extensive documentation, comprehensive tests of the package's internal functions, and common use cases are provided, sufficiently covering the current OpenAlex API. The source code and development versions are available at https://github.com/ropensci/openalexR. The current version of the package is a stable version, and there are no plans for any breaking changes soon. Of course, **openalexR** will continue to be actively maintained to keep up with CRAN policies and distribute any bug fixes. Bug reports, help requests, or improvement suggestions are welcome in the package software repository. For more information on **openalexR**, vignettes are available at https://ropensci.github.io/openalexR/articles/.

7 Acknowledgements

8 Supplementary material

Examples we show in this manuscript can be found at https://github.com/trangdata/oarj/blob/main/paper-examples.md.

References

W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, et al. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*, 2018. [p167]

M. Aria. *openalexR: Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API*, 2022. https://github.com/massimoaria/openalexR, https://massimoaria.github.io/openalexR/. [p169]

M. Aria and C. Cuccurullo. bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of informetrics*, 11(4):959–975, 2017. [p169]

A. Belfiore, A. Salatino, and F. Osborne. Characterising research areas in the field of ai. *arXiv preprint arXiv:2205.13471*, 2022. [p168]

D. S. Chawla. Unpaywall finds free versions of paywalled papers. *Nature*, 2017. [p168]

C. Chen. Science mapping: A systematic review of the literature. *Journal of Data and Information Science*, 2(2):1–40, 2017. doi: doi:10.1515/jdis-2017-0006. URL https://doi.org/10.1515/jdis-2017-0006. [p167]

G. Csárdi, J. Hester, H. Wickham, W. Chang, M. Morgan, and D. Tenenbaum. *remotes: R Package Installation from Remote Repositories, Including 'GitHub'*, 2021. URL https://CRAN.R-project.org/package=remotes. R package version 2.4.2. [p170]

T. Dallas, A.-L. Gehman, and M. J. Farrell. Variable bibliographic database access could limit reproducibility. *BioScience*, 68(8):552–553, 2018. [p167]

C. Du, J. Cohoon, J. Priem, H. Piwowar, C. Meyer, and J. Howison. Citeas: Better software through sociotechnical change for better software citation. *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, 2021. [p168]

G. Hendricks, D. Tkaczyk, J. Lin, and P. Feeney. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427, 2020. [p167]

C. Herzog, D. Hook, and S. Konkiel. Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1):387–395, 2020. [p167]

D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, and I. Rafols. Bibliometrics: the leiden manifesto for research metrics. *Nature*, 520(7548):429–431, 2015. [p167]

D. W. Hook, S. J. Porter, and C. Herzog. Dimensions: building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3:23, 2018. [p167]

P. Kulkanjanapiban and T. Silwattananusarn. Comparative analysis of dimensions and scopus bibliographic data sources: an approach to university research productivity. *International Journal of Electrical & Computer Engineering (2088-8708)*, 12(1), 2022. [p167]

A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E. Delgado López-Cózar. Google scholar, microsoft academic, scopus, dimensions, web of science, and opencitations' coci: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1):871–906, 2021. [p167]

S. McWeeny, J. Choe, and E. S. Norton. *SnowGlobe: An Iterative Search Tool for Systematic Reviews and Meta-Analyses*, 2021. [p169]

S. McWeeny, S. Choi, J. Choe, A. LaTourrette, M. Y. Roberts, and E. S. Norton. Rapid automatized naming (ran) as a kindergarten predictor of future reading in english: A systematic review and meta-analysis. *Reading Research Quarterly*, 57(4):1187–1211, 2022. doi: https://doi.org/10.1002/rrq.467. URL https://ila.onlinelibrary.wiley.com/doi/abs/10.1002/rrq.467. [p169]

T. L. Pedersen. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*, 2022a. URL https://CRAN.R-project.org/package=ggraph. R package version 2.1.0. [p175]

T. L. Pedersen. *tidygraph: A Tidy API for Graph Manipulation*, 2022b. URL https://CRAN.R-project.org/package=tidygraph. R package version 1.2.2. [p169, 175]

J. Priem, D. Taraborelli, P. Groth, and C. Neylon. Altmetrics: A manifesto. 2011. [p167]

J. Priem, H. Piwowar, and R. Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022. URL http://altmetrics.org/manifesto. [p168]

A. P. Siddaway, A. M. Wood, and L. V. Hedges. How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, 70(1):747–770, 2019. doi: 10.1146/annurev-psych-010418-102803. [p169]

V. K. Singh, P. Singh, M. Karmakar, J. Leta, and P. Mayr. The journal coverage of web of science, scopus and dimensions: A comparative analysis. *Scientometrics*, 126(6):5113–5142, 2021. [p167]

A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246, 2015. [p167]

N. J. Van Eck, L. Waltman, V. Larivière, and C. Sugimoto. Crossref as a new source of citation data: A comparison with web of science and scopus. *CWTS Blog*, 17, 2018. [p167]

R. Van Noorden. Scientists join journal editors to fight impact-factor abuse. *Nature News Blog*, 16, 2013. [p167]

M. Visser, N. J. van Eck, and L. Waltman. Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic. *Quantitative Science Studies*, 2 (1):20–41, 2021. [p167]

K. Wais. Gender prediction methods based on first names with genderizer. *R J.*, 8(1):17, 2016. [p169]

L. Waltman and V. Larivière. Special issue on bibliographic data sources, 2020. [p167]

K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, 2020. [p167]

K. Wang et al. A review of microsoft academic services for science of science studies. front. big data 2, 45 (2019), 2019. [p167]

S. B. Wanyama, R. W. McQuaid, and M. Kittler. Where you search determines what you find: the effects of bibliographic databases on systematic reviews. *International Journal of Social Research Methodology*, 25(3):409–422, 2022. [p167]

H. Wickham. *httr: Tools for Working with URLs and HTTP*, 2022. URL https://CRAN.R-project.org/package=httr. R package version 1.4.4. [p169]

H. Wickham, J. Hester, W. Chang, and J. Bryan. *devtools: Tools to Make Developing R Packages Easier*, 2022. URL https://CRAN.R-project.org/package=devtools. R package version 2.4.5. [p170]

D. J. Winter. rentrez: An r package for the ncbi eutils api. Technical report, PeerJ Preprints, 2017. [p167]

C. Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, EASE '14, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450324762. doi: 10.1145/2601248.2601268. URL https://doi.org/10.1145/2601248.2601268. [p169]

Massimo Aria
Università degli Studi di Napoli Federico II
K-Synth srl, Academic Spin-off
Department of Economics and Statistics
Napoli, NA 80126
Italy
(0000-0002-8517-9411)
aria@unina.it

Trang Le
Bristol Myers Squibb
Cambridge, MA 02143
USA
(0000-0003-3737-6565)
trang.le@bms.com

Corrado Cuccurullo
Università della Campania Luigi Vanvitelli
Capua, CE 81043
Italy
Università degli Studi di Napoli Federico II
K-Synth srl, Academic Spin-off
Department of Economics and Statistics
Napoli, NA 80126
Italy
(0000-0002-7401-8575)
corrado.cuccurullo@unicampania.it

Alessandra Belfiore
Università degli Studi di Napoli Federico II

K-Synth srl, Academic Spin-off
Department of Economics and Statistics
Napoli, NA 80126
Italy
(0000-0003-3709-9481)
alessandra.belfiore@unina.it

June Choe
University of Pennsylvania
Philadelphia, PE 19104
USA
(0000-0002-0701-921X)
yjchoe@sas.upenn.edu