# Using OpenAlex to Analyse Cited Reference Patterns

Eric Schares\* and Sandra Mierz\*\*

\**eschares@iastate.edu*

ORCID 0000-0002-6292-8221

University Library, Iowa State University, United States

*\*\* sandra.mierz@proton.me*

ORCID 0000-0002-8913-9011

## Abstract

Understanding what material an institution cites is an important piece of journal renewal or cancellation decisions. Many proprietary products exist that can provide this type of data but require a paid subscription to access. Therefore, this project develops an open and repeatable process to extract cited reference data using OpenAlex, a freely available database of publication metadata.

The code is written in Python and publicly provided in two Jupyter Notebooks. Part 1 demonstrates how to use the OpenAlex API to extract publications which meet user-defined criteria and collect the cited references within. Part 2 provides a standardized set of graphs, data visualizations, and tables to explore and answer questions about cited reference patterns. A case study of one university is provided.

Using an open dataset lets anyone run this analysis, and posting the code publicly makes the procedure reproducible, saves time, and allows for extensibility by others.

## 1. Introduction

Collection development decisions in an academic library require an understanding of the institution’s overall usage of a journal title. When deciding to renew or cancel a particular title, download and publication numbers are relatively straightforward to gather but understanding citation usage is more difficult. What material do authors at a research institution use when writing their publications? These outgoing citations may be to material from a wide range of disciplines stretching back any number of years, resulting in large and complex cited reference datasets.

Various tools exist that can analyze reference information, but these are typically part of a paid database and require a user to be authenticated beyond a paywall to gain access. An overview of many of these tools is provided in Waltman (2016). Examples of proprietary products that are capable of citation analysis include Web of Science (Birkle et al., 2020), InCites (Clarivate, 2023), SciVal (Elsevier, 2023), and Dimensions (Digital Science, 2020; Herzog et al., 2020).

By contrast, the database OpenAlex (Priem et al., 2022, https://openalex.org/) is a freely available Knowledge Graph which includes cited reference data as one of its connections among publications, researchers, and datasets. The data is openly accessible via an API and has a CC0 license attached, which allows users to analyse and reuse it.

In this project, we demonstrate the use of the OpenAlex API to collect, clean, and examine one year’s worth of open cited reference data from Iowa State University-authored publications. This process is automated using Python and Jupyter Notebooks and will address the following research questions:

RQ1. How many references are cited in total and per publication?

RQ2. What journals and publishers are cited, and how often?

RQ3. What years were the cited articles published?

RQ4. What is the average reference age per publication?

## 2. Method

This section describes how to use the OpenAlex API to find all journal articles published by Iowa State University authors in 2021 and gather their references. The accompanying Python code can be found in the Jupyter Notebook "Part 1" (Schares & Mierz, 2023).

*2.1. Data source*

OpenAlex is a fully-open scientific knowledge graph (Priem et al., 2022, https://openalex.org/)), that contains metadata of around 250M scholarly works including their citations and references (OpenAlex, 2023a). Each work is identified by an internal identifier, the OpenAlex ID, which is used to connect it to other works. For example, every work has an attribute “referenced\_works” in its metadata (OpenAlex, 2023b) that consists of a list of OpenAlex IDs to identify its references.

The primary way to query data from OpenAlex is via its API. Each entity type (works, authors, institutions, sources, publishers, concepts) has its own endpoint that can be queried using a variety of filters to specify a desired subset.

*2.2. Data collection*

To obtain the data for this project, we used the OpenAlex API via the “works” endpoint on January 27, 2023 to query for all publications by Iowa State University authors published in 2021 and collected the cited references listed in the bibliographies.

To get all publications, we filtered for works having at least one authorship affiliation with Iowa State University that were published between January 1, 2021 and December 31, 2021 and were classified as journal articles, but not as paratext (cover, table of contents, issue information). Since the OpenAlex API paginates its result set, we used cursor paging to get the complete results.

The references of each publication are specified in their “referenced\_works” attribute.

After downloading the publications, we extracted the OpenAlex IDs from all the “referenced\_works” attributes into a single list and removed duplicate entries to avoid redundant API calls. Following the approach outlined in the OurResearch blog (Meyer, 2022) we used one API call for each slice of 50 OpenAlex IDs from the list to request all references.

*2.3. Data storage*

Once the data is retrieved, it is stored into local files to avoid repeatedly requesting the same information. This not only saves time but also reduces the burden on the OpenAlex API.

To minimize the amount of data storage needed, only a subset of attributes relevant to the data analysis is extracted from the metadata of both publications and references and preserved. These attributes include id, doi, title, publication\_year, host\_venue.display\_name, host\_venue .publisher, host\_venue .issn\_l and number of referenced\_works.

Following the design for many-to-many relationships in relational databases, the data for publications, references and their connections are stored separately (Monge, 2014). Publications are stored in a human readable .csv file format. Because the number of references is much higher than the number of publications, we switched to the .parquet format to store references, which uses compression to save on data storage but is not human readable.

The connections between publications and their references are stored in a dedicated .csv file called "pub2ref.csv". Each entry has the form of a tuple of OpenAlex IDs, one identifying the publication, the other one if its references. This file serves as a so-called join table and allows us to efficiently join the publications with their corresponding references into one united dataset (Figure 1).

Figure 1: Connections between publications and references through a join table “pub2ref”



With the data collection and storage complete, the exploration of patterns and investigating the characteristics of the cited references can begin.

## 3. Results

The result for this example is 3,350 articles from Iowa State University authors published in 2021 and their 142,961 cited references. The Juypter Notebook “Part 2” provides a series of pre-made graphs and data visualizations that help analyse citation patterns and make exploring the citation data easy to understand.

*3.1. Number of references*

In this dataset, 394 publications report no cited references (11.8%), but manual investigation shows the articles do have some. Citation data is not openly available for every publication, and OpenAlex can only return data it is aware of. The Initiative for Open Citations (I4OC, 2022; Peroni & Shotton, 2020) continues to advocate for publishers to make their citation information openly available, and while there has been success, the data is not 100% complete.

The 394 publications that report 0 references can be removed from the dataset, but the opposite problem emerges – one large outlier of a publication with 4,075 references skews the distribution to the right. Figure 2 shows the cumulative distribution of number of references in the 2,956 publications with at least one reference. This is zoomed in and does not show 8 publications (0.2%) which have over 300 cited references.

The 50% point of this distribution is 40 cited references, meaning half of the publications have 40 references or fewer, and half have 41 references or more. The high reference count publications skew the mean higher, to 48.3 references per article. For comparison, an analysis of author behaviour in the German Projekt DEAL showed a median number of 32 references per article (Fraser et al., 2023).

Figure 2: Cumulative distribution of number of references



*3.2. Journal titles and publisher*

Over 11,500 unique journals were cited by Iowa State University authors in 2021. Table 1 shows the top 10 journals by times cited in the dataset.

Table 1: Top 10 journals by number of cited references

|  |  |  |
| --- | --- | --- |
| **Title** | **Publisher** | **Citations** |
| Proc. of Nat’l Acad. of Sciences | National Academy of Sciences | 1804 |
| Physical Review | American Physical Society | 1655 |
| Science | AAAS | 1535 |
| Nature | Nature Portfolio | 1457 |
| PLOS ONE | Public Library of Science | 1326 |
| Physical Review Letters | American Physical Society | 1214 |
| Jnl of the American Chemical Society | American Chemical Society | 999 |
| Nature Communications | Nature Portfolio | 968 |
| Scientific Reports | Nature Portfolio | 912 |
| Journal of Animal Science | American Society of Animal Science | 688 |

Figure 3 expands this list to show the top 50 titles grouped by publisher. Nature Portfolio publishes three of the top ten most cited journals in this dataset (Table 1), and five of the top 50 (Figure 3, purple). The American Physical Society publishes two of the top ten, and three of the top 50 (blue). Elsevier publishes five of the top 50 (green), but none of the top 10 most cited journals. Each of the five Elsevier journals in the top 50 have between 300 and 550 citations in this dataset, in contrast to Nature and APS which have a smaller number of very highly cited journals.

Figure 3: Top 50 journals by number of cited references, coloured by publisher



Combining individual journal titles into overall publisher counts, Elsevier is the most cited publisher in this dataset, its 31,103 cited references accounting for 21.7% of the data (Figure 4). This is 2.9x more than the second-place publisher, Wiley-Blackwell, which has 10,764 cited references and accounts for 7.5% of the total.

Figure 4: Total number of citations by publisher.



We can also track one journal of interest, for example *Physical Review Letters*, and see most references are to material from 2020-1996 (Figure 5).

Figure 5. Year of material cited in *Physical Review Letters*.



*3.3. Article characteristics and date of reference*

123,381 unique articles were cited in this dataset. The publication referenced most often was cited 59 times (Perdew et al., 1996). Citation activity extends back to 1690, with the oldest citation to “An Essay concerning Human Understanding” by John Locke (1690). Figure 6(a) shows a histogram of the year a cited reference was published, with 2019 highlighted as the single year most often cited in this dataset (7.4%). The x-axis of Figure 6(a) extends back to cover 1690 and the single publication cited in that year.

Figure 6. Histogram of year published, showing full distribution (a) and zoomed in (b).

 

(a) (b)

Figure 6(b) shows the same histogram but is zoomed in to make it easier to see. Somewhat surprisingly, there is material published in 2021 that was already cited in 2021 (left-most bar), with the rate roughly equal to material published in 2003. After the peak of activity in 2019, each year moving back in time is cited less often than the more recent year, with a small exception in the years 1981 and 1982. Even in situations where the bars appear to be equal (such as 2018 and 2017), close inspection does reveal that 2017 is slightly below 2018, at 7.15% to 7.14%. These distributions follow previous studies on citation activity, with peaks two to three years prior to publication and subsequent slow declines (Fraser et al., 2023; Parolo et al., 2015).

To compare citation patterns of publications, one approach could be to draw multiple curves on the plot; however, with over 3,000 total publications this would quickly become difficult to read and compare. Therefore, we summarize this data by finding the average age of each paper’s cited references. Table 2 shows the average year delta of the youngest and oldest reference lists, calculated by taking the year of publication (2021, in this example) and subtracting the year of each reference, then averaging those year deltas together to find one summary number per publication.

Table 2. The five smallest and five largest average year deltas.



Somewhat surprisingly, eight publications in this dataset have an average year delta of 0, meaning they only cite material from the year they were published! Looking more closely shows several of these cite themselves through a Suggested Citation that is picked up by OpenAlex, while the Correction and Introductions refer to specific publications from the same year.

The cumulative distribution plot of average year deltas for all publications is shown in Figure 7. Once again, there is a large outlier that skews the plotting of this curve far to the right – a publication with 23 cited references and an average year delta of 138 years. This paper’s topic, “North American species of Rubus L. (Rosaceae) described from European botanical gardens (1789-1823)” fits with its references skewing older than typical.

Figure 7. Cumulative distribution of average year delta



To learn more, we can track this publication in more detail by separating it from the main distribution. Figure 8 breaks out the 23 references from this oldest article and turns them red, showing the citation patterns from that single paper.

Figure 8. Distribution of cited references from paper with largest average age, broken out from group and highlighted in red.



## 4. Discussion

The number of cited references per paper ranged from 0 to over 4,000. When the 0 reference papers were removed, the median number was 40 per paper and the mean 48. This shows how one paper with a great many (excessive?) references pulls up the mean. The year most often cited in this dataset was 2019. The median age of a reference was 8 years old, and the mean was 11.4 years. Half the references in this set were to material published between 2021 and 2013, with the other half ranging from 2013-1690. Each publication has its own average reference age, and the median of this factor was 11 years.

Even though this analysis is based on citation count, it should be noted that citation count is not the best or only way to assess journal titles or articles. Subject speciality needs to be included when weighing collection decisions for renewal or cancellation. While we have provided many data points and graphs to assist in understanding cited reference patterns, there is still no substitute for human knowledge and subject expertise on journal collection decisions.

Overall, there are an almost infinite number of questions and plots one could make with a dataset as rich as this. The project and code we have provided is intended to help users at other research organizations first collect their cited reference data, and then get them started answering common questions. Perhaps new and unusual patterns will come to light, which will then start a new round of investigation with a slightly different focus, and so on. With this foundation in place, the limit is only what an interested researcher can think up to investigate.

## 5. Limitations and Conclusion

OpenAlex is a very useful open scholarly database, but it is still evolving. As such, executing the notebooks at different points in time may yield different results, and bugs may appear which were not present in January 2023 when this data was downloaded and processed. To be as transparent as possible, we will list problems that arise when executing the notebooks as we become aware of them on the project’s GitHub repository (Schares & Mierz, 2023).

We have demonstrated how to understand the characteristics of what is cited by researchers at an institution. Other universities and research institutions can run similar analyses using the freely available code, and we welcome and encourage this use. Simply editing the ROR ID in the “Part 1” notebook will change which research institution is being investigated, and the provided graphs and data visualizations will update accordingly. The range of dates can be adjusted to modify which time range is under investigation. This work is intended to help libraries worldwide better understand the citation patterns of researchers at their institution, reduce duplication of effort, and speed up the creation of knowledge.

## Open Science Practices (100-200 words)

This project is completely reproducible and open to being extended and built upon. We have used OpenAlex, a freely available open data source, and the data and code from the project are publicly available at <https://github.com/eschares/OpenAlex-CitedReferences> (Schares & Mierz, 2023).

In fact, this collaboration and conference paper is itself is a direct result of Open Science practices. A preliminary version of this project was delivered at the Workshop on Open Citations & Open Scholarly Metadata 2022. In that short presentation, Schares demonstrated a functional, but limited, proof of concept. The code was posted openly on GitHub where Mierz discovered it and made many substantial improvements to the functionality and reduced the execution time. As a result, the data collection was expanded from one day of publications to an entire year. While we have never met, our successful international collaboration is a testament to the principles and possibilities of Open Science.

## Acknowledgements

The authors wish to thank the team at OurResearch for developing and supporting OpenAlex and making such rich data freely available. We also wish to thank the organizers of STI 2023 and any reviewers that improve the manuscript.

This paper was written using data obtained from the OpenAlex API on January 27, 2023. Plots were created using Plotly version 5.13.0 (Plotly, 2022).

**Author contributions**

E.S. coded the initial attempt to collect references in OpenAlex, created data visualizations, and wrote the paper.

S.M. rewrote and improved the code regarding the data download from OpenAlex (reducing run time by 20x and handling datafile sizes), better organized and standardized the GitHub repository, and wrote the methods section of this paper.

**Competing interests**

The authors declare no competing interests.

**Funding information**

The authors declare no funding was received for this research.

**References**

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies, 1*(1), 363-376. <https://doi.org/10.1162/qss_a_00018>

Clarivate. (2023). *InCites Benchmarking & Analytics*. <https://clarivate.com/products/scientific-and-academic-research/research-analytics-evaluation-and-management-solutions/incites-benchmarking-analytics>

Digital Science. (2020). *Citation Analysis: Journals Cited by a Research Organization*. <https://api-lab.dimensions.ai/cookbooks/2-publications/Which-Are-the-Journals-Cited-By-My-Organization.html>

Elsevier. (2023). *SciVal*. <https://www.elsevier.com/solutions/scival>

Fraser, N., Hobert, A., Jahn, N., Mayr, P., & Peters, I. (2023). No Deal: German Researchers’ Publishing and Citing Behaviours after Big Deal Negotiations with Elsevier. *Quantitative Science Studies*, 1-33. <https://doi.org/10.1162/qss_a_00255>

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies, 1*(1), 387-395. <https://doi.org/10.1162/qss_a_00020>

I4OC. (2022). *Initiative for Open Citations*. <https://i4oc.org/>

Locke, J. (1690). *An essay concerning human understanding* (K. P. Winkler, Ed.). Hackett.

Meyer, C. (2022). *Fetch multiple DOIs in one OpenAlex API request*. <https://blog.ourresearch.org/fetch-multiple-dois-in-one-openalex-api-request/>

Monge, A. (2014). *Database design with UML and SQL* (4th ed.). Wiley.

OpenAlex. (2023a). *Counts*. Retrieved 2023-04-20 from <https://api.openalex.org/counts>

OpenAlex. (2023b). *Work object: Referenced works*. <https://docs.openalex.org/api-entities/works/work-object#referenced_works>

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics, 9*(4), 734-745. [https://doi.org/https://doi.org/10.1016/j.joi.2015.07.006](https://doi.org/https%3A//doi.org/10.1016/j.joi.2015.07.006)

Perdew, J. P., Burke, K., & Ernzerhof, M. (1996). Generalized Gradient Approximation Made Simple. *Physical Review Letters, 77*(18), 3865-3868. <https://doi.org/10.1103/physrevlett.77.3865>

Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies, 1*(1), 428-444. <https://doi.org/10.1162/qss_a_00023>

Plotly. (2022). *Plotly open source graphing library for Python [software]*. <https://plotly.com/python>

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*.

Schares, E., & Mierz, S. (2023). *OpenAlex-CitedReferences*. <https://github.com/eschares/OpenAlex-CitedReferences>

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391. <https://doi.org/10.1016/j.joi.2016.02.007>