From manual to automated:
Enhancing publication tracking at the California Academy of Sciences (CAS) using the OpenAlex API

Comprehensive publication records for biodiversity collections can highlight their significance. While collecting all published specimen citations is challenging, author affiliation metadata facilitates gathering research linked to an institution. We use author affiliations and ORCIDs to build an institutional publication record from OpenAlex search results, which include papers, datasets, and other academic works. This method improves attribution for CAS researchers compared to previous ad hoc methods.

Maricela Abarca
California Academy of Sciences
University of San Francisco
0000-0002-0890-8887
mabarca@calacademy.org

Joseph Russack
California Academy of Sciences
0000-0001-8366-0941
jrussack@calacademy.org

Problem

In 2019, Texas A&M spent about $212,000 on Web of Science and $140,000 on Scopus, two popular publication tracking services.[1]

These high costs force CAS to manually track publications. While inexpensive, this approach is time-consuming, prone to errors, and often overlooks early-career researchers, students, and non-curators.

Solution

Combine OpenAlex results with a list of CAS researchers to create an affordable publication database for reports and metrics.

OpenAlex is an open-source, global catalog of research, authors, and institutions. It rivals Web of Science and Scopus but is freely available (CC0 license) and user-friendly. It includes works in other languages and is committed to representing research from the Global South.[2]

Project Design

Docker Container

OpenAlex + python

Collect and organize data from OpenAlex API calls.

Powered by MySQL

Ingest reorganized data into a MySQL database.

Generate csv reports via the command line.

Interested in trying this for your institution? Visit Github for docs and examples: https://github.com/calacademy-research/publications_finder

Select Findings

2022 Total Publication Count

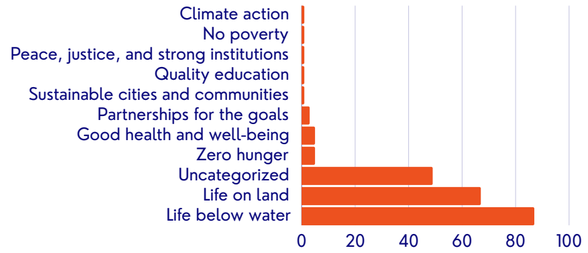
CAS manual tracking vs. OpenAlex

66 vs **199**

66 CAS publications appeared in a 2022 internal report, about one-third of the 199 OpenAlex results for that year. (Excluding published datasets).

Data as of August 8, 2024

2022 Publication Count by UN Sustainability Goal



OpenAlex's Sustainability Goal labels for CAS works align with CAS expertise (e.g. terrestrial and marine collections and research). However, the algorithm left 49 CAS works uncategorized in 2022. Results include published datasets.

Data as of May 22, 2024

2,111

The number of times the **CAS Mammalogy GBIF dataset** has been cited. This is the **3rd most cited work** in the CAS OpenAlex publication profile.

Data as of August 8, 2024

Takeaways

OpenAlex is easy to use and well documented for both technical and non-technical users.
- The web interface facilitates searches and filtering if you want a simpler approach than ours.

Searching by institution results in better attribution than manual tracking.
- This can help reduce reliance on rosters of affiliated researchers, which are often hard to maintain.
- Searching with ORCIDs augments results.

Vet OpenAlex results for "overmerging": when the disambiguation algorithm combines distinct authors under the same research profile.
- You can submit data curation requests when these errors are found.

Future Directions

Continue to monitor API query results.
- Share the results with stakeholders.
- Add a form to submit publications to our database that are missing from OpenAlex.

Set up an interactive public dashboard.
- Integrate other metrics and visualizations, including specimen citations.

References 1 Web of Science versus Scopus: Journal Coverage Overlap Analysis, Simona Tabacaru, April 2019, accessed August 7, 2024. https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/175137/Web%20of%20Science%20versus%20Scopus%20Report%202019.pdf?sequence=4&isAllowed=y

2 Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833

Thank you to Newmark Women in Tech and SPNHC for supporting conference travel costs.