جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology | University Library

Exporting Institutional Repository Metadata as Dataset

Marcelo Garcia Daryl Grenz
Mohamed Ba-Essa

University Library, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Motivation

Datasets play a crucial role in scientific research, and it is expected of researchers to publish their datasets. Although we generally consider datasets to be the outcome of an experiment or a sensor, a university library also holds a dataset. This dataset is comprised of the university's publications. Interestingly, it seems that institutional repositories don't publish their metadata as a bulk dataset, possibly on the assumption that making the information available for harvesting through interfaces such as OAI-PMH is sufficient to support reuse.

We think a separate bulk dataset in a commonly used format may be useful to potential users who are not able or interested in working through OAI-PMH. For example, such a dataset could be used by administrators of the university for reporting purposes, or by students who need a sample of thousands of files to train their models. Finally, having a list of the university's outputs in an easy to use format may have uses we can't imagine yet.

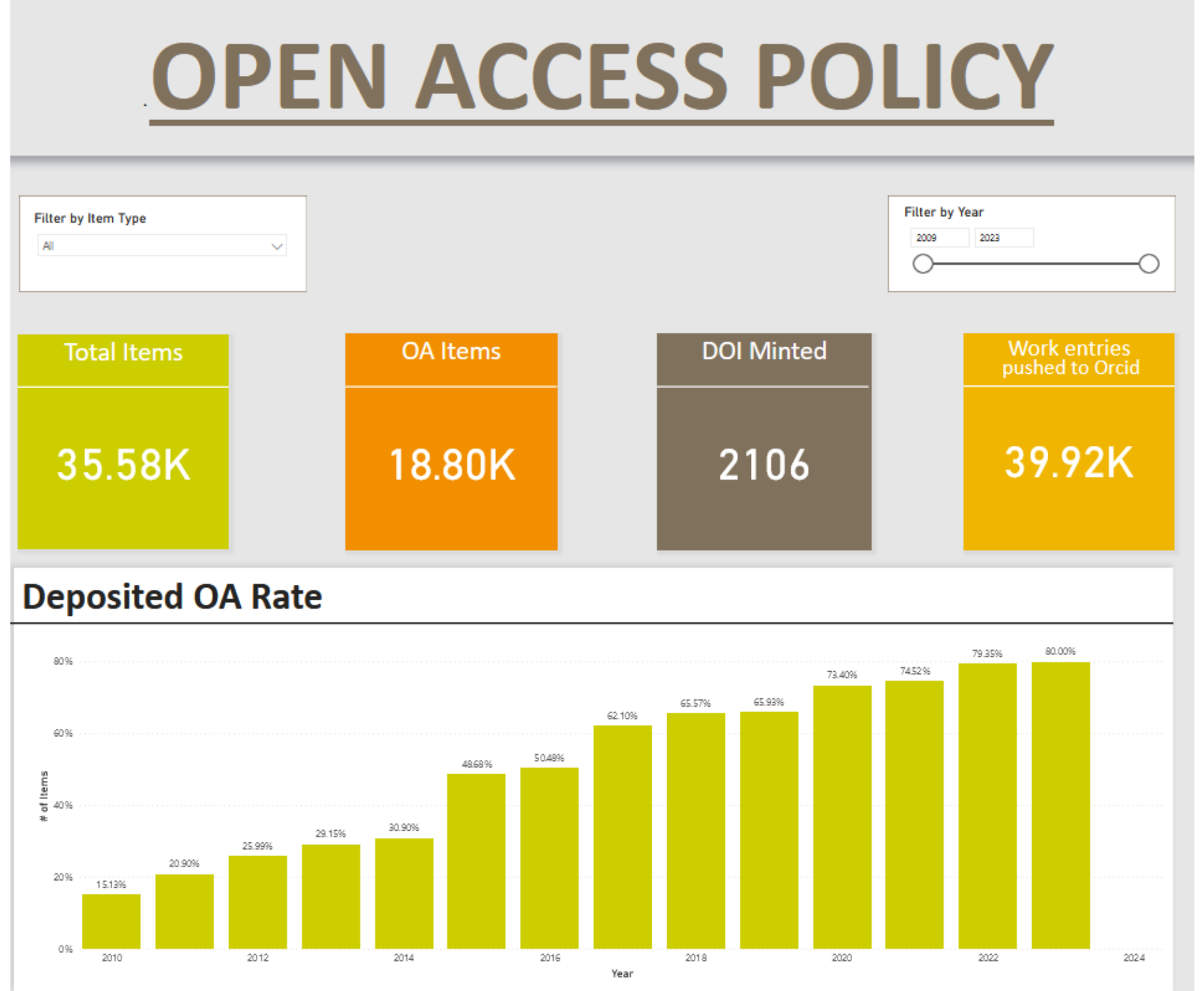
The inspiration for this work came from the CORD-19 project lead by the Allen Institute for AI (AI2). Where they aggregated information about research papers related to COVID-19 and published an archive of these papers in a machine-readable format so the AI community could use them.

Based on this example, we decided to explore the idea of exporting the content of our institutional repository as a dataset. The outcome is a CSV file, a format that is easy to use in data science or AI workflows.

Methodology

We query our local publication database for selected fields and use the result to generate a CSV file that is publicly posted in a dataset record in our institutional repository. We chose fields that are common to many kinds of outputs, like author, abstract, publication date, and type of publication. We also include a link to an output's full text PDF (if available) and to its extracted text to facilitate the process of tokenization. At this stage we didn't include the text directly in the dataset. It was necessary to change the separator of the authors field to make easier to export to VOSviewer.

Suggested Use



OPEN ACCESS POLICY

Filter by Item Type: All

Filter by Year: 2009 – 2023

Total Items	OA Items	DOI Minted	Work entries pushed to Orcid
35.58K	18.80K	2106	39.92K

Deposited OA Rate

Figure 1. Open Access Dashboard

The idea of this dataset was to invite the KAUST community to explore the academic output of the university. We hope that our community will find other uses, and come back with feedback, for example about which other fields should be included. We already have some ideas of potential uses:

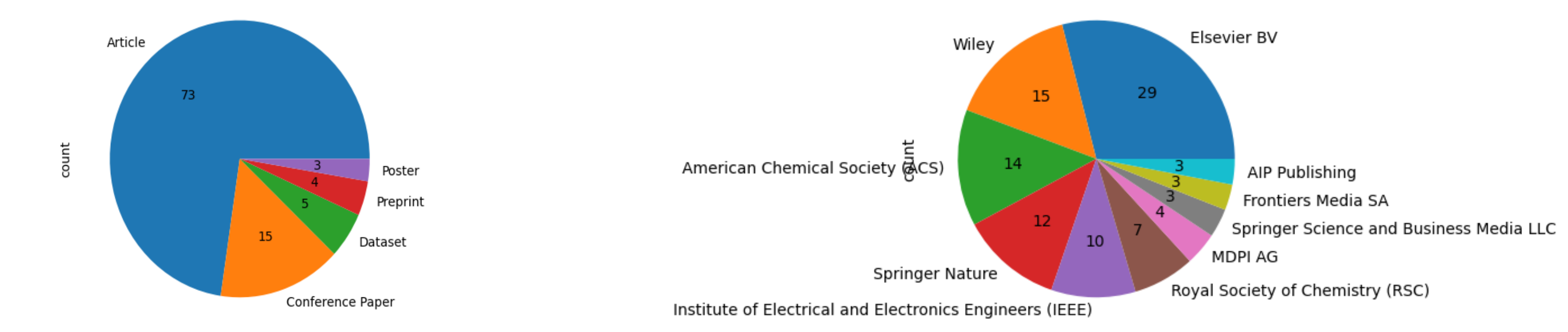
- As the data source for the library's open access dashboard (1). This will require additional open access and departmental information be included.
- To allow users to quickly get a copy of metadata for thousands of KAUST affiliated research outputs that they can sort or filter as they desire without being limited by the search or export functionality of the repository software.
- Allow users to quickly assess the overlap between this list and lists in other databases, such as Scopus, Web of Science or Lens.org by including all of the known external ids in the dataset.

Usage

As an example we created a Jupyter notebook as part of a Github repository, the repository includes the notebook itself and a sample CSV file for development. The full dataset is available on KAUST's institutional repository. The links are below:

- Dataset: `https://repository.kaust.edu.sa/handle/10754/691065`
- Notebook: `https://github.com/kaust-library/RepoDataset`

Using the full dataset, we make simple queries like the percentage of articles, presentations, preprints, etc. See Fig (2a). Or, from the articles, the percentage across the top 10 publishers. See Fig (2b)



(a) Types of publications

(b) Top Publishers

Figure 2. Types and Publishers

Another example of usage would be to load into an exploration tool like VOSviewer, to see the relation between elements. We used one of the notebook in GitHub to create a graph in VOSviewer of the elements in the field *abstract*, see Fig (3).

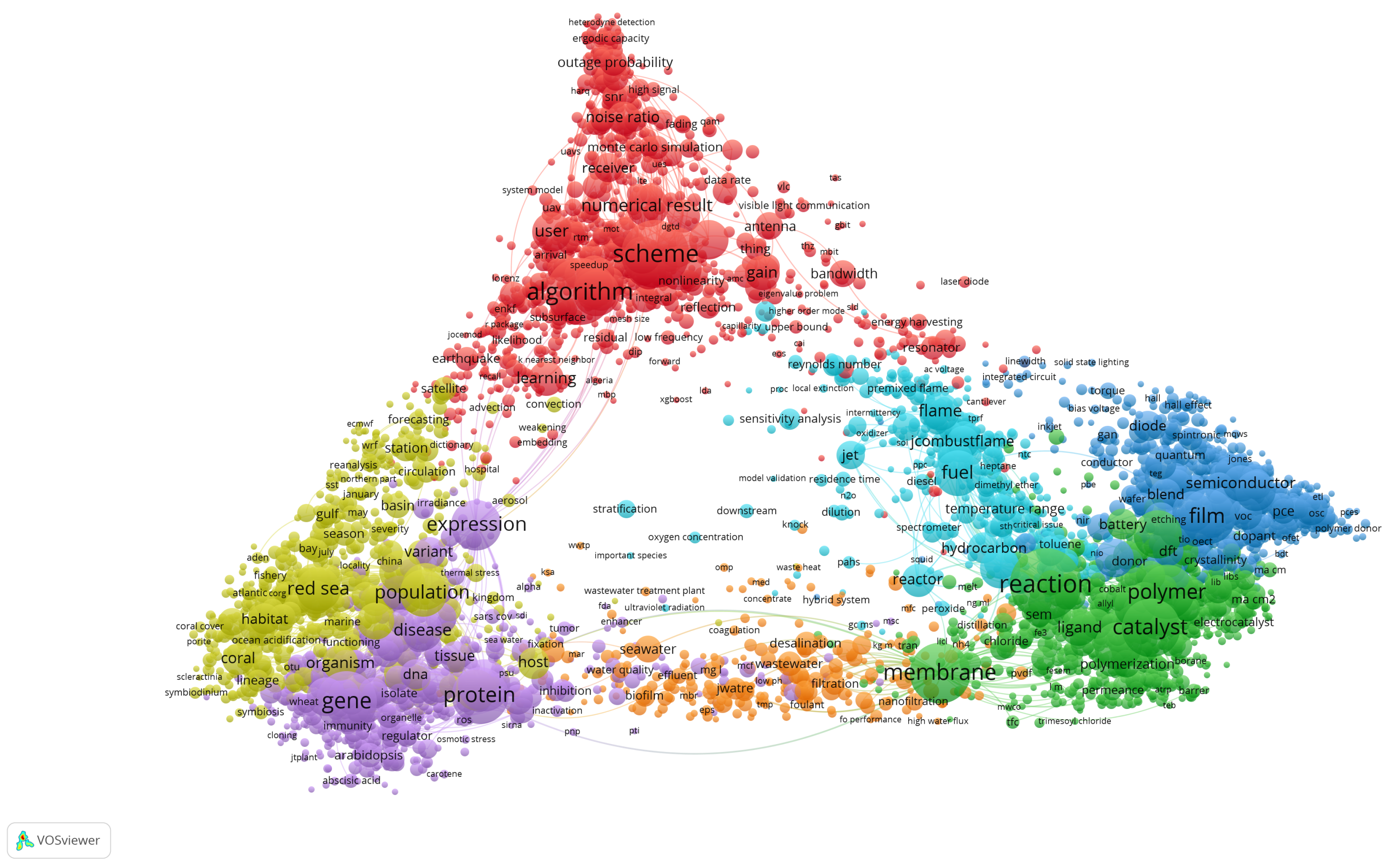


Figure 3. Visualization of items in the abstract.

Conclusion

The idea of this project was to explore another venue for the library to engage with our community, and more specifically with data scientists and AI practitioners. We hope the community will use the dataset and provide feedback on how they are using the dataset and how we can improve the service.

References

[1] Nees Jan van Eck and Ludo Waltman.
VOSviewer visualizing scientific landscape.
`https://www.vosviewer.com/`.

[2] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier.
CORD-19: The COVID-19 open research dataset.
In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.