Bibliometric Crawling Framework & Digital Identifier Analysis

Asura Enkhbayar, Know-Center (Austria)
aenkhbayar@know-center.at

Motivation

In a scientific publishing environment that is increasingly moving online, unique and persistent identifiers of scholarly work are gaining in importance.
We investigated the distribution and coverage of identifiers from articles using different data sources (*arXiv, CrossRef, Mendeley and the SAO/NASA ADS*[a]).

The initial study [2] was conducted with a small dataset consisting of ∼14,000 articles from the discipline of quantitative biology (arXiv short code: *q-bio*).

[a] Smithsonian Astrophysical Observatory (SAO)/NASA Astrophysics Data System (ADS)

Data and Software

arXiv Online pre-print repository that is mainly self-archived by authors

CrossRef Official DOI[a] registration agency

Mendeley Reference Manager and social network for researchers. Publications can be uploaded to Mendeley

SAO/NASA ADS Developed by NASA; contains mostly astronomy and physics papers

In order to access these sources mainly the official APIs were accessed directly, whereas in the cases of arXiv and CrossRef existing R-packages were incorporated.[b]

[a] Digital Object Identifier
[b] aRxiv, rcrossref - both from ropensci.org

Crawling Framework

Written in Python & R. Two operating modes have been implemented:

arXiv→CrossRef→Mendeley arXiv disciplines are crawled subdiscipline-wise. These dataframes are then enhanced with additional DOIs from CrossRef. Finally Mendeley is queried with either arxiv-id or DOI.

arXiv→SAO/NASA ADS arXiv disciplines are the starting point again, but afterwards SAO/NASA ADS

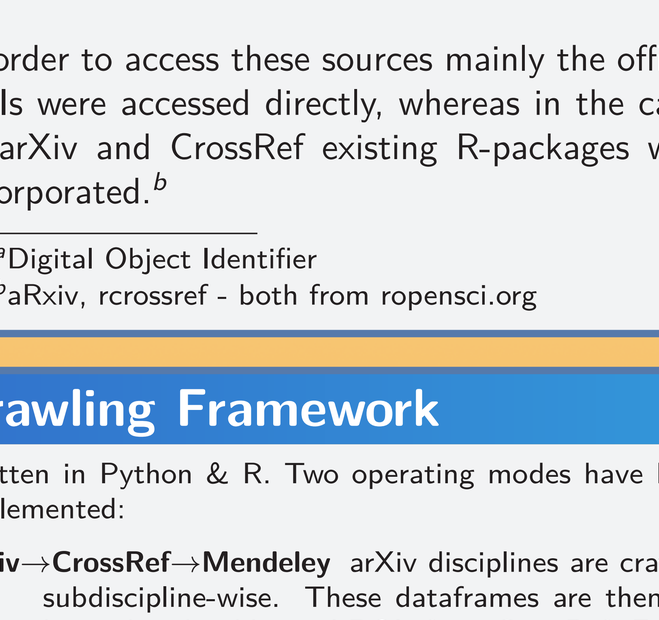

Figure 1: Flowchart of the arXiv-CrossRef-Mendeley collection pipeline

https://github.com/Bubblbu/crawling-framework

Analysis of Bibliometric Identifiers

Initial study

We found that when retrieving arXiv articles in quantitative biology from Mendeley, we were able to obtain more articles using the DOI than the arXiv ID. Even when we only considered articles that were assigned both identifiers, the effect was sizeable (91.4% vs. 72.6%). This indicates that the DOI may be a better identifier with respect to findability. Nevertheless, a single arXiv ID is the second most popular identifier combination on Mendeley. This suggests that pre-prints are being read — if at a lower level — even when they are not yet published in a journal.

We found that coverage of articles on Mendeley decreases in the most recent years, whereas the availability of DOIs does not decrease in the same order of magnitude. This hints at the fact that there is a certain time lag before articles are covered in crowd-sourced services on the scholarly web.

Multi-disciplinary study

Ensuing from the first study we have collected a bigger dataset and are currently investigating the behaviour of identifiers among different disciplines (*physics*, *cs*, *math* and *q-bio*)

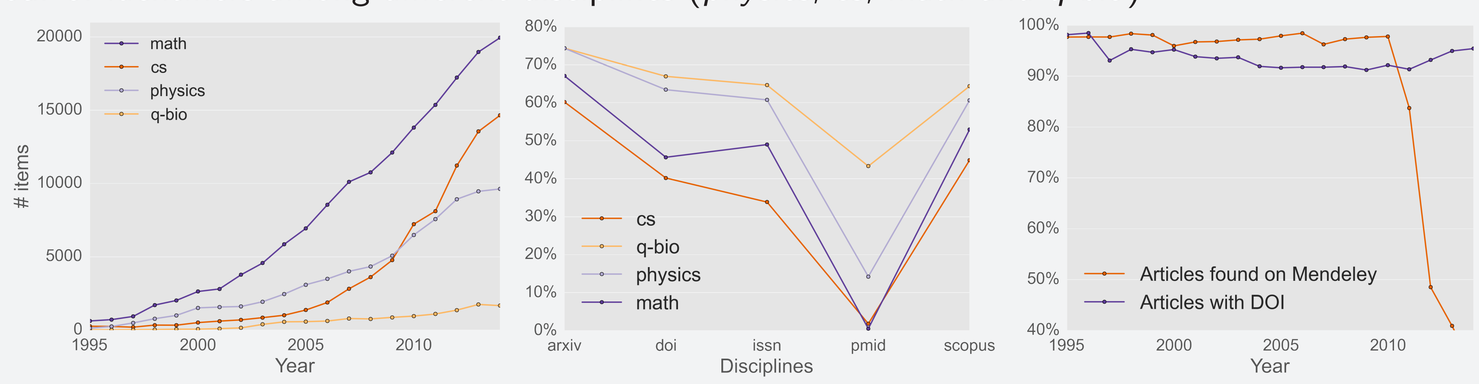

Figure 2: (1) Growth of articles among disciplines, (2) Identifier distribution among disciplines, (3) Findability of articles on Mendeley vs DOI coverage

Headstart

Head Start is intended to help researchers that are new to a field with their literature search, utilising co-readership patterns found in Mendeley. The interactive visualisation presents the main areas in the field and lets you zoom into relevant publications. See Kraker et al. [3]

Visualisation of SAO/NASA ADS Data

Using the second crawling pipeline of the framework we collected multiple ADS datasets from different disciplines. As ADS does not provide the same co-readership information as Mendeley, alternative similarity measures and clustering algorithms were tested.

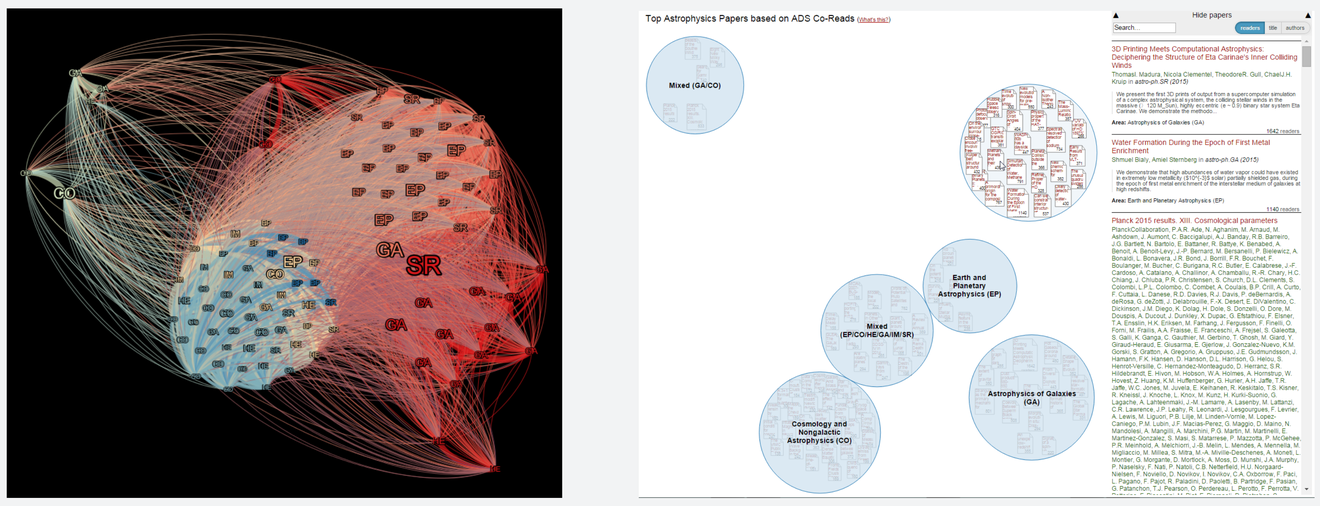

Figure 3: (1) Graph-based clustering and visualisation of 300 astrophysics articles inspired by Bollen et al. [1]; (2) Same 300 articles ported to Headstart

Outlook

The analysis of all arXiv entries over the full span of disciplines is very enticing, especially considering that arXiv has celebrated its one millionth article in January 2015. With further development of the framework this amount of articles should pose no problem for regular analyses.

At the moment the crawling framework partially relies on external packages, whereas native support of the APIs would be preferable because the maintenance and quality of packages fluctuates over time.

References

[1] Johan Bollen et al. "Clickstream Data Yields High-Resolution Maps of Science". In: *PLoS ONE* 4.3 (Mar. 2009), e4803. DOI: 10.1371/journal.pone.0004803. URL: http://dx.doi.org/10.1371/journal.pone.0004803.

[2] Peter Kraker, Asura Enkhbayar, and Elisabeth Lex. "Exploring Coverage and Distribution of Identifiers on the Scholarly Web". In: *Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science, ISI 2015, Zadar, Croatia, May 19-21, 2015.* 2015, pp. 393–403.

[3] Peter Kraker et al. "Visualization of co-readership patterns from an online reference management system". In: *Journal of Informetrics* 9.1 (2015), pp. 169–182. ISSN: 1751-1577. DOI: http://dx.doi.org/10.1016/j.joi.2014.12.003. URL: http://www.sciencedirect.com/science/article/pii/S1751157714001151.