Exploring Coverage and Distribution of Identifiers on the Scholarly Web

Peter Kraker[1], Asura Enkhbayar[1], Elisabeth Lex[2]

[1] Know-Center
Inffeldgasse 13/VI, 8010 Graz, Austria
{pkraker, aenkhbayar}@know-center.at
[2] Graz University of Technology
Inffeldgasse 13/V, 8010 Graz, Austria
elisabeth.lex@tugraz.at

Abstract

In a scientific publishing environment that is increasingly moving online, identifiers of scholarly work are gaining in importance. In this paper, we analysed identifier distribution and coverage of articles from the discipline of quantitative biology using arXiv, Mendeley and CrossRef as data sources. The results show that when retrieving arXiv articles from Mendeley, we were able to find more papers using the DOI than the arXiv ID. This indicates that DOI may be a better identifier with respect to findability. We also find that coverage of articles on Mendeley decreases in the most recent years, whereas the coverage of DOIs does not decrease in the same order of magnitude. This hints at the fact that there is a certain time lag involved, before articles are covered in crowd-sourced services on the scholarly web.

Keywords: Scholarly identifiers, Pre-prints, arXiv, DOI, Readership

In: F. Pehar/C. Schlögl/C. Wolff (Eds.). Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19th–21st May 2015. Glückstadt: Verlag Werner Hülsbusch, pp. 393–403.

1 Introduction

In a scientific publishing environment that is increasingly moving online, identifiers of scholarly work are gaining in importance. With the advent of pre-print archives, there is often more than one version of an article available and these versions may be hosted in various places around the web. Scholarly communication is no longer limited to articles alone, but it also takes place in different forms on various social media platforms. Identifiers are therefore crucial for disambiguation and traceability of scholarly articles and their reception.

The need for persistent identifiers is often mentioned in the literature (see e.g. Davidson & Douglas, 1998; Bourne & Fink 2008) and consequently, a variety of identifier systems have been proposed (see e.g. Van De Sompel et al., 2001; Warner 2010). Prominent examples for identifiers on an article level are the Digital Object Identifier or DOI (DOI Foundation, n.d.) and the arXiv ID. Notable identifiers on the author level are author-based identifiers such as ORCID (Haak et al., 2012) and Researcher ID (Thomson-Reuters, n.d.). Some of the most longstanding identifiers predate the digital age, including the International Standard Book Number (ISBN) and the International Standard Serial Number (ISSN).

Despite their importance, little is empirically known about the coverage and distribution of scholarly identifiers, and how they propagate on the scholarly web. In our work, we are addressing this very gap in the scientometric literature. Specifically, our research was guided by the following research questions:

- How are scholarly identifiers distributed in crowd-sourced systems, e.g. pre-print archives and online reference management systems? Which identifier combinations are the most common? Who are the top providers of identifiers?
- Does the provision of different identifiers have an influence on findability of scientific publications in other bibliographic and bibliometric sources?

2 Data and method

In this study, we analysed arXiv papers from the discipline of quantitative biology (arXiv short code: q-bio). We chose this discipline because it represents one of the largest disciplines on Mendeley (Kraker et al., 2012). Three different data sources were used in this study: (i) arXiv, a preprint archive (ii) CrossRef, a metadata and linking service, and (iii) Mendeley, an online reference management system.

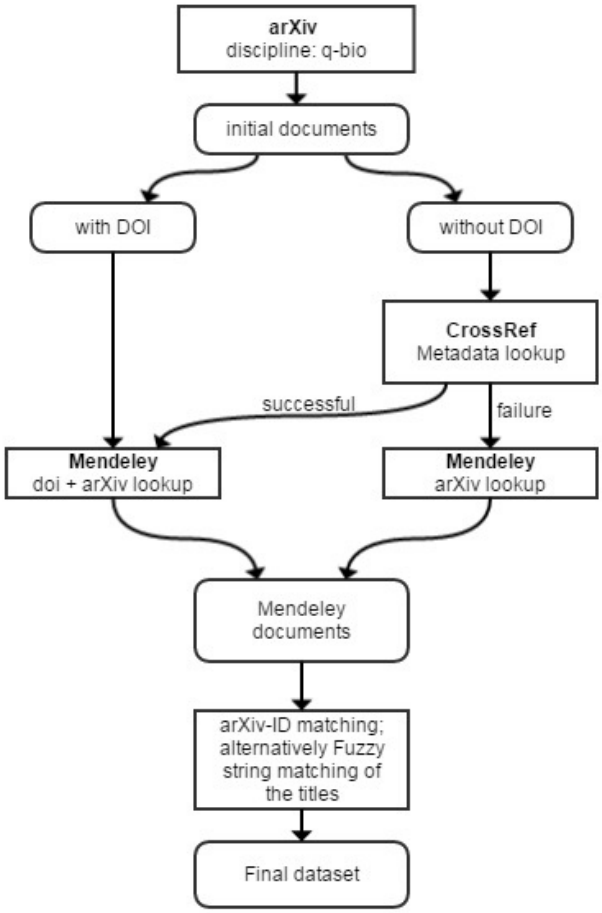

Figure 1. Data collection pipeline

The data collection pipeline is shown in figure 1. At first, we collected metadata on all publicly available articles for quantitative biology. In all cases, the most recent upload to arXiv was used and all older entries were discarded. This resulted in n = 14,195 metadata records. Quantitative biology represents a medium-to-small collection on arXiv. The collected metadata includes: arXiv ID, DOI (optional), title, authors, year, and journal (optional).

This data was sourced on 17.11.2014 and was used as a basis for all following steps. At first, the initial data set was divided into entries with DOI (n = 5,125 entries, 36.7%) and without a DOI (n = 8,980 entries, 63.3%). arXiv is primarily used as a way to disseminate pre-prints, and not all authors add a DOI to the arXiv record after an article has been published. Therefore, we performed a CrossRef meta-data lookup in order to acquire additional DOIs. We used the following metadata to search for an entry: title, author, journal, and year.

With this procedure, we found DOIs for an additional 1,885 entries, bringing the number of entries with a DOI up to 7,100 (50.02%). We then attempted to retrieve the corresponding documents for all entries on Mendeley. We used either the arXiv ID or both the DOI and the arXiv ID to locate the document. If both arXiv ID and DOI yielded a result on Mendeley, the Mendeley IDs were compared. If they didn't match, we used the result, which contained additional identifier fields, e.g. a PubMed ID, if available. If both results contained the same amount of articles, we chose the item found with the DOI.

Finally, we compared the arXiv ID of the obtained Mendeley document with the original arXiv entry. If the obtained Mendeley document did not provide one, the two titles were compared using approximate string matching in order to ascertain matching documents.

After this procedure, we arrived a final set of n = 11,570 articles that could be found on Mendeley (81.5%). For these articles, we retrieved basic readership data and identifier data. Available identifiers on Mendeley are:[1]

- arxiv: arXiv ID
- doi: Digital Object Identifier (DOI)
- isbn: International Standard Book Number (ISBN)
- issn: International Standard Serial Number (ISSN)
- pmid: PubMed ID (assigned to publications indexed in PubMed)

1 See http://dev.mendeley.com/methods/#catalog-documents.

- scopus: Scopus ID (assigned to publications indexed in Scopus)
- ssrn: Social Science Research Network (SSRN) ID

3 Results

3.1 Identifier distribution in arXiv and findability on Mendeley

Table 1 sums up the basic results of the crawling process. Of the 14,195 unique articles, 36.7% had a DOI on arXiv. Using CrossRef, an additional 1,885 DOIs could be found, bringing the share of articles with a DOI up to 50.02%. 11,570 articles (81.5%) could finally be found on Mendeley.

Table 1. Results of the crawling process; n = 14,195 articles

arXiv: total docs	arXiv: docs with DOI	CrossRef: additional DOIs	Mendeley: found
14,195	5,125 (36.7%)	1,885 (13.3%)	11,570 (81.5%)

There was a difference in findability with respect to whether we used a DOI or the arXiv ID to search for the articles on Mendeley (see also table 3). Of the 14,195 articles, 72.6% could be retrieved on Mendeley using the arXiv ID. In contrast to that, 91.4% of the 7,100 articles with a Digital Object Identifier (either on arXiv or via metadata lookup on CrossRef) could be found on Mendeley using the DOI.

One of the reasons for that could be that records with a DOI do represent articles that have eventually been published in a journal. In order to test this assumption, we analysed the registrants for all entries with a DOI (7,100 articles). We used a list of DOI registrants by Alf Eaton[2] with manual extensions to identify registrants. The results confirm our assumption (see table 2). The top registrants are established publishers such as Elsevier and Springer. These publishers usually assign DOIs to articles published in their journals and books, in contrast to archives such as figshare, which assign a DOI to any submitted article regardless of whether it was published in a journal or not.

Table 2. DOI registrants of articles; n = 7,100 articles

2 See https://gist.github.com/hubgit/5974843.

Registrant	# DOIs	Percentage
American Physical Society	1,507	21.2%
Elsevier	1,029	14.5%
Springer-Verlag	668	9.4%
Public Library of Science	502	7.1%
IOP Publishing	439	6.2%
American Institute of Physics	335	4.7%
Proceedings of the National Academy of Sciences	217	3.1%
Oxford University Press	194	2.7%
Springer (Biomed Central Ltd.)	180	2.5%
IOP Publishing – Europhysics Letters	141	2.0%
Other	1,888	26.6%
Sum	7,100	100%

To eliminate effects that relate to the nature of the article that has been posted on arXiv (whether it stayed a pre-print or went on to become a journal article), we also compared findability for articles that have both a DOI and an arXiv ID (see table 3). We also found a difference in these cases: 91.4% of articles with a DOI could be found using the very same identifier, whereas, only 71.4% of articles with a DOI could be found with the arXiv ID. The lowest findability was reported for articles with no DOI: of the 7,095 articles with no DOI, only 69.0% were retrieved using the arXiv ID.

Table 3. Findability of articles on Mendeley, depending on the identifier used; n = 14,195 articles

		found on Mendeley using	
	n	arXiv ID	DOI
arXiv ID & DOI	7,100 (50.02%)	5,414 (76.25%)	6,492 (91.44%)
arXiv ID	7,095 (49.98%)	4,896 (69.01%)	-
Sum	14,195 (100%)	10,310 (72.63%)	-

Another interesting fact found in the top providers is that the American Physical Society, which is, among other things, "working to advance and diffuse the knowledge of physics through its outstanding research journals"[3]

3 See http://www.aps.org/about/index.cfm.

is the top registrant for DOIs in quantitative biology. One of the reasons for that could be that arXiv allows authors to assign more than just one category to each article. The analysis of article categories (see table 4) shows that quantitative biology is the primary discipline for only 61.4% of articles with a DOI (4,358 articles). 30.1% (2,178 articles) are assigned to a primary category that falls into the discipline of physics. This indicates a high number of interdisciplinary articles in the sample.

Table 4. Distribution of disciplines in articles with a DOI (n = 7,100 articles)

Discipline	Number of articles	Percentage
Quantitative Biology	4,358	61.4%
Physics	2,178	30.7%
Computer Science	247	3.5%
Mathematics	211	3.0%
Statistics	105	1.5%
Quantitative Finance	1	0.0%
All	7,100	100.0%

Figure 2 shows the distribution of articles from 1992 to 2013. There is a strong, at times exponential increase in the number of articles. The coverage on Mendeley, however, has declined for the youngest articles as can be seen in figure 3. The percentage of articles with a DOI does not decrease in the same order of magnitude.

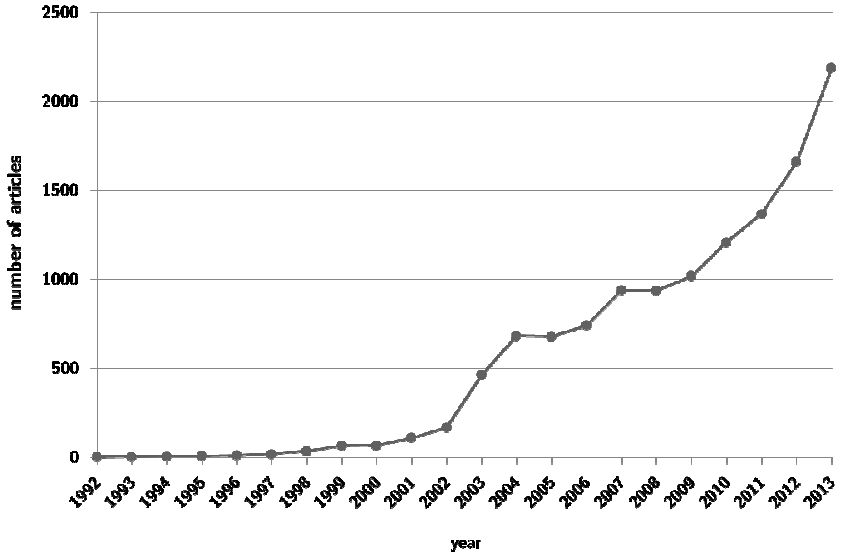

Figure 2. Distribution of articles between 1992 and 2013; n = 12,392 articles

Figure 3. Findability of articles on Mendeley and DOI coverage, 1992–2013; n = 12,392 articles

3.2 Distribution of identifiers on Mendeley

We then investigated the distribution of identifiers of all arXiv articles found on Mendeley in detail. Note that we only took metadata from Mendeley into account, which is why the numbers for arXiv ID and DOI differ to the analyses before. The distribution of identifiers on Mendeley can be seen in table 5. The arXiv ID is the most common identifier, followed by the Scopus ID, DOI and ISSN. In terms of readership, articles with a PubMed ID have the highest average readership.[4]

Table 5.
Identifier frequency and mean readership on Mendeley; n = 11,570 articles

	arxiv	doi	scopus	pmid	issn
frequency	10,351 (89.5%)	8,321 (71.9%)	8,409 (72.7%)	5,477 (47.3%)	8,119 (70.2%)
mean readership	20.4	25.4	25.4	32.4	25.9

Figure 4 shows the most common identifier combinations in the data. Here, a combination of all identifiers on Mendeley included in this analysis (arXiv ID, DOI, ISSN, PubMed ID and Scopus ID) is the most common identifier combination; a single arXiv ID comes second.

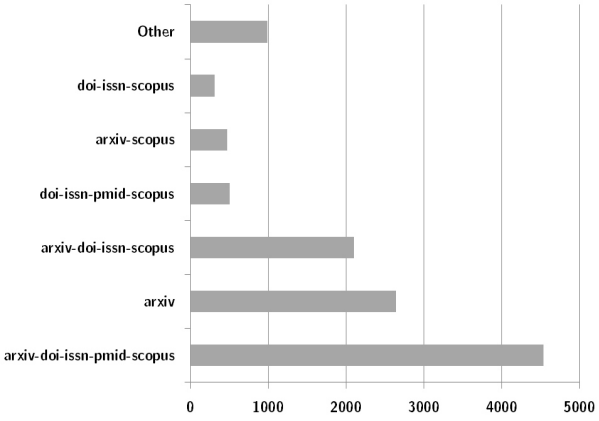

Figure 4.
Identifier combination frequency of articles on Mendeley; n = 11,570 articles

4 Note that we left ISBN out of this analysis, because the metadata quality was very poor with respect to this field on Mendeley.

4 Conclusions and future work

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.

There are certain limitations to this work. We only looked at a single discipline (quantitative biology) and we only used three data sources in our study (arXiv, CrossRef and Mendeley), which may have had a significant influence on the results. Indeed, in a small-scale study using a random sample of 381 articles from Web of Science, Zahedi et al. (2014) report that they were able to retrieve only 47.7% of articles on Mendeley using the DOI or the title.

In the future, we therefore plan to extend this study to more disciplines and fields in order to substantiate the hypotheses emanating from the results in this study. In order to gain a deeper insight into the distribution and the coverage of identifiers on the scientific web, we are looking to include further data sources such as Web of Science, PubMed Central, Altmetric.com, and figshare.

Acknowledgments

The Know-Center is funded within the Austrian COMET program – Competence Centers for Excellent Technologies – under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

Bourne, P. E., & Fink, J. L. (2008). I am not a scientist, I am a number. *PLoS Computational Biology*, 4 (12), e1000247. doi:10.1371/journal.pcbi.1000247

Davidson, L. A., & Douglas, K. (1998). Digital Object Identifiers: Promise and Problems for Scholarly Publishing. *The Journal of Electronic Publishing*, 4 (2). doi:10.3998/3336451.0004.203

DOI Foundation (n.d.). Digital Object Identifier System. http://www.doi.org/ <14.01.2015>.

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, 25 (4), 259–264. doi:10.1087/20120404

Kraker, P., Körner, C., Jack, K., & Granitzer, M. (2012). Harnessing User Library Statistics for Research Evaluation and Knowledge Domain Visualization. In: *Proceedings of the 21st International Conference Companion on World Wide Web*, pp. 1017–1024. doi:10.1145/2187980.2188236

Sompel, H. Van De, Lagoze, C., Bekaert, J., Liu, X., Payette, S., & Warner, S. (2006). An interoperable fabric for scholarly value chains. *D-Lib Magazine*, 12 (10).

Thomson-Reuters (n.d.). ResearcherID. URL: http://www.researcherid.com <14.01.2015>.

Van de Sompel, H., & Beit-Arie, O. (2001). Open linking in the scholarly information environment using the OpenURL Framework. *New Review of Information Networking*, 7 (1), 59–76. doi:10.1080/13614570109516969

Warner, S. (2010). Author identifiers in scholarly repositories. *Journal of Digital Information*, 11 (1), 1–10.

Zahedi, Z., Haustein, S., & Bowman, T. D. (2014). Exploring data quality and retrieval strategies for Mendeley reader counts. In: *Metrics14 – ASIS&T Workshop on Informetric and Scientometric Research Introduction*. http://www.asis.org/ SIG/SIGMET/data/uploads/sigmet2014/zahedi.pdf <04.03.2014>.

Mapping the Spreading of Cited References over Research Fronts of Bibliographically Coupled Publications

Edgar Schiebel

AIT Austrian Institute of Technology GmbH,
Donau City Straße 1, 1220 Wien, Austria
edgar.schiebel@ait.ac.at

Abstract

This short paper deals with the delineation of research issues based on bibliographic coupling of publications assisted by visualization techniques. Cluster techniques, multidimensional scaling or spring models reveal agglomerations of similar publications but it is always difficult to have a clear picture of the thematic substructure of a research field or even a set of publications with a consistent content. The central research questions of this work are: How can we visualize the occurrence of cited references in an agglomeration of similar publications? Does the visualization of the occurrence of cited reference in bibliographically coupled publications help to understand how to delineate a research topic? Research fronts were defined as a local agglomeration of similar publications in a two dimensional space. This work proposes a visualization method using an overlay technique in 2D heat maps of bibliographically coupled publications. With this approach we could visualize and discuss to what extend research fronts are formed by several highly cited references that are the core of the underlying knowledge base. The approach is demonstrated for research related to foresight.

Keywords: Bibliographic coupling, Science mapping, Visualization, Overlay technique, Delineation of research issues

In: F. Pehar/C. Schlögl/C. Wolff (Eds.). Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19th–21st May 2015. Glückstadt: Verlag Werner Hülsbusch, pp. 404–409.