Preserving Scholarly Communications on the Web with Open Metadata

Martin Czygan[1], Nathaniel Smith[1]

[1] *Internet Archive*

Introduction

Scholarly communication documents on the web, particularly in the long tail, are exposed to at least two kinds of decay: they vanish entirely (Laakso, Matthias, & Jahn, 2021) or they are affected by citation or reference rot (Klein et al., 2014). Since 2017, the Internet Archive has increased its focus on preserving scholarly material on the web to fulfill its mission, and also to address data decay issues. The access site at scholar.archive.org (https://scholar.archive.org) was launched in March 2021; the first iteration of a citation graph, called refcat was released in October 2021 (Czygan, Holzmann, & Newbold, 2021)[1]. In this paper, we highlight a few aspects of the preservation process and recent developments, specifically bulk bibliographic datasets and how potentially relevant links are discovered, as well as provide a brief update on the citation graph dataset (v3).

Open Bibliographic Metadata

First, preservation of scholarly communications on the web relies on openly available metadata. In the past, we have included metadata gathered from bibliographic aggregators like PubMed (Canese & Weis, 2013), DBLP (Ley, 2002), and DOAJ (Morrison, 2017) and from DOI registrars like CrossRef (Hendricks, Tkaczyk, Lin, & Feeney, 2020) and DataCite (Brase, 2009) into our continuous cataloging pipeline. In addition, we used data from the (now discontinued) Microsoft Academic (Sinha et al., 2015), which has been carried forward in OpenAlex (Priem, Piwowar, & Orr, 2022), for targeted web archiving. We also expanded our metadata acquisition from the web by accessing more than 150,000 endpoints across over 40,000 domains implementing OAI-PMH (Lagoze, Van de Sompel, Nelson, & Warner, 2002), an application-level protocol used by popular web publishing tools like Open Journal Systems (Willinsky, 2005) and institutional repository software, such as DSpace, among others. A recent version of this dataset, which we call oaiscrape[2], included about 200M[3] metadata records in the dublin core format (Weibel, Kunze, Lagoze, & Wolf, 1998) representing various types of records (publications, theses, datasets, media and others). After a cleanup process, we discovered over 300M unique URL candidates[4] in this dataset alone (across all metadata

[1] Most bibliographic data artifacts are available under the Bulk Bibliographic Metadata collection (https://archive.org/details/ia_biblio_metadata).

[2] We make this dataset available in regular intervals as part of a metadata collection at IA biblio metadata (https://archive.org/details/ia_biblio_metadata)

[3] We observe 196204847 unique records by identifier. While most records have distinct identifiers, there exist records in the dataset (and upstream data) sharing an identifier, while referring to different records, which is permitted by the standard (Lagoze et al., 2002).

[4] For an exact count additional rounds of data reviews are required.

fields[5]); these 300M+ URL are distributed across about 280K different domains[6]. Table 1 shows the frequency of the top-10 domains referenced in the oaiscrape records. Among them, we also find over 600K links to git hosting sites (mainly github.com), suggesting mentions of software projects, project websites, or datasets (Escamilla et al., 2023).

Table 1. A list of the top-10 domains of links found in the oaiscrape dataset (2025)

Count	Domain
30535604	hdl.handle.net
15740750	doi.org
14469950	pt.cision.com
8604306	gallica.bnf.fr
7127858	www.kb.dk
6791854	figshare.com
5599077	prensahistorica.mcu.es
5224331	kb-images.kb.dk
4757477	dx.doi.org
3915844	hal.science

From these URL candidates, we generate seed lists for targeted crawls, which we conduct with Heritrix (Mohr, Stack, Rnitovic, Avery, & Kimpton, 2004). Crawlers use custom configurations for the task, addressing different web crawling situations such as persistent identifier redirects (up to a maximum), following Google Scholar meta-tags[7], and others.

Estimation of Metadata Overlap in Large Bibliographic Datasets

Internet Archive Scholar uses data from different sources to guide archiving efforts, and a regular question is: to what extent does the metadata of these different sources overlap? We ran an analysis across seven bibliographic data sources to understand the amount of duplication in the metadata[8] - we looked at overlaps based on DOI in CrossRef, DataCite, DOAJ, DBLP, oaiscrape, OpenAlex and PubMed. When we concatenate all DOIs present in these seven datasets, we find 451,395,081 persistent identifiers. When keeping only unique

[5] While the OAI DC XML schema has more preferred fields for a url, e.g. dc:identifier, not all metadata providers follow the guidelines strictly.
[6] After verification against the current List of Top-Level Domains from ICANN: https://www.icann.org/resources/pages/tlds-2012-02-25-en
[7] As described on: Google Scholar: Inclusion Guidelines for Webmasters
[8] Analysis conducted in 01/2025

identifiers, we arrive at 235,399,926 DOIs[9]. We load the DOI lists into a DuckDB database (Raasveldt & Mu ̈hleisen, 2019) and calculate the cardinality of their pairwise intersections and differences, of which the results can be found in Table 2.

Table 2. Pairwise comparison of DOI in bulk upstream datasets (01/2025). We extract and normalize the DOI found in the various datasets, then calculate their intersection and differences. All datasets pairs have DOI in common.

A	**B**	**card(A)**	**card(B)**	**A ∩ B**	**A \ B**	**B \ A**
crossref	datacite	165,644,551	62,966,529	177,878	165,466,673	59,394,867
crossref	dblp	165,644,551	6,461,206	6,028,963	159,615,588	432,243
crossref	doaj	165,644,551	9,004,954	7,411,626	158,232,925	1,593,328
crossref	oaiscrape	165,644,551	6,794,290	4,926,119	160,718,432	1,868,171
crossref	openalex	165,644,551	169,057,060	157,182,783	8,461,768	11,874,277
crossref	pubmed	165,644,551	31,466,491	26,771,400	138,873,151	4,695,084
datacite	dblp	62,966,529	6,461,206	303,452	59,269,293	6,157,754
datacite	doaj	62,966,529	9,004,954	107,795	59,464,950	8,897,159
datacite	oaiscrape	62,966,529	6,794,290	905,195	58,667,550	5,889,095
datacite	openalex	62,966,529	169,057,060	8,060,828	51,511,917	160,996,232
datacite	pubmed	62,966,529	31,466,491	10,566	59,562,179	31,455,918
dblp	doaj	6,461,206	9,004,954	284,159	6,177,047	8,720,795
dblp	oaiscrape	6,461,206	6,794,290	258,262	6,202,944	6,536,028
dblp	openalex	6,461,206	169,057,060	6,344,982	116,224	162,712,078
dblp	pubmed	6,461,206	31,466,491	355,755	6,105,451	31,110,729
doaj	oaiscrape	9,004,954	6,794,290	509,456	8,495,498	6,284,834
doaj	openalex	9,004,954	169,057,060	7,737,253	1,267,701	161,319,807
doaj	pubmed	9,004,954	31,466,491	3,830,635	5,174,319	27,635,849
oaiscrape	openalex	6,794,290	169,057,060	4,556,644	2,237,646	164,500,416
oaiscrape	pubmed	6,794,290	31,466,491	662,073	6,132,217	30,804,411
openalex	pubmed	169,057,060	31,466,491	26,968,174	142,088,886	4,498,310

[9] According to DOI.org approximately 300M DOI have been assigned to date

We find that each source has unique DOI contributions (see Table 3) validating our approach of using multiple upstream sources to build a comprehensive metadata catalog.

Table 3. Unique DOI contributions from the analyzed datasets (01/2025); number of DOI we find in a dataset, but in none of the others.

Dataset	Unique contributions (DOI)
datacite	50,689,363
crossref	7,764,417
pubmed	4,263,905
openalex	3,362,058
doaj	1,006,264
oaiscrape	820,600
dblp	63,537

Additional Link Discovery with Sitemaps

At scale, data acquisition with the OAI-PMH mechanism can exhibit some challenges, as endpoints use a wide variety of implementations on the server side, from well tested, widely used open source projects to ad-hoc implementations – sometimes not even emitting well-formed XML. In general, we are interested in the bibliographic data, and harvesting OAI-PMH metadata offers a computationally lightweight way to acquire this data, as opposed to more elaborate methods, such as analysis of PDF data with GROBID (Romary & Lopez, 2015)[10] or similar tools.

Additionally, metadata quality varies considerably across endpoints, which usually requires additional code to compensate[11]. When focusing on preservation, we likely care about publications first, and metadata second.

Websites may choose to implement sitemaps (Sitemaps XML format, 2005). Sitemaps follow a standard schema and can use plain text or XML to list a number of links on a given domain, in addition to metadata such as date of last update. While the sitemap format is standardized, their URLs is not, although common locations exist. We iteratively discovered the sitemap

[10] For GROBID processing, we typically use multicore machines; as GROBID can utilize deep learning models for article segmentation, it benefits from the presence of a GPU. Meanwhile, metadata harvesting can be done on low-end systems, merely having enough storage space available.

[11] Typical issues include incomplete data, multiple data items in a single field, duplication, inconsistent formatting, among many other issues.

locations specifically for domains running Open Journal Systems (OJS)[12] and expanded them, if necessary[13].

We limited the sitemap exploration to OJS, because OJS contains typically only few pages not directly related to actual publications (and we are interested in targeted crawls). We generated a list of candidate sitemap locations and then used a high-performance link check tool[14] to confirm existence. A minority of the assumed sitemap locations were valid; however we were able to generate a list of about 40M raw URLs and then trim this list down to 19M likely HTML or PDF links, due to the regular URL structure of sites implementing OJS. After deduplicating the 19M URL against already preserved holdings at the Internet Archive we were able to generate a crawl seedlist of 10,081,479 unseen URL from this approach[15].

Internet Archive Scholar Citation Graph Update

In February 2024, we released an updated version (v3) of our Internet Archive Scholar citation graph, called refcat[16], for short. After starting with over 3.5B raw reference entities collected from metadata and from running GROBID (Romary & Lopez, 2015) over archived PDF documents, we employed a series of matching techniques to detect relations to publications in our catalog. The most recent version contains 2.173B edges, an increase of about 64% compared to the initial dataset from 2021. While most edges are again determined by exact matches (i.e. by identifier), about 7% of the edges are found through various fuzzy matching algorithms developed at the Internet Archive[17]. In the spirit of the Open Citations project (Peroni & Shotton, 2020) we aim to increase the amount of publicly available reference data to foster open science and reuse of this kind of data for a variety of applications. As our next iteration's catalog will likely see an increase in metadata coverage, and we expect the next iteration of the citation graph to be more comprehensive as well, as references are matched against this catalog.

Summary and Outlook

We aim to continuously build and extend our metadata catalog, while providing access through scholar.archive.org and making various data artifacts and open source software tools available in the process. With the extension of our catalog, we expect the derived citation graph to become more comprehensive over time as well. We also aim to further extend the discovery of potential scholarly material by applying metadata and link-gathering techniques – similar to the described approach using sitemaps – to other software tools regularly found in use for institutional repositories and open access journals. The significant increase in usage of large language models (Minaee et al., 2024) in recent years prompted a number of projects aiming to extract structure from binary documents formats, among them docling (Team, 2024)

---

[12] Via: https://openjournaltheme.com/what-is-the-ojs-omp-sitemap-location/

[13] We used a helper CLI tool called sitemapped. A sitemap can contain direct links or links to other sitemaps. For example, the single sitemap for the CORE project found under core.ac.uk/sitemap.xml would expand to over 200M links.

[14] Many tools in this space exist, we used a link checker called clinker, which can run several hundred checks in parallel.

[15] The corresponding crawl was conducted between 07/2024 and 01/2025 yielding 4.1TB of crawl data.

[16] Available under https://archive.org/details/refcat_2024-02-15

[17] Code available at https://gitlab.com/internetarchive/refcat

and markitdown (markitdown, 2024), and we are evaluating these tools for our indexing pipeline, which feeds documents to the search engine underlying our access portal.

References

Brase, J. (2009). Datacite - a global registration agency for research data. In 2009 fourth international conference on cooperation and promotion of information resources in science and technology (pp. 257–261).

Canese, K., & Weis, S. (2013). Pubmed: the bibliographic database. The NCBI handbook, 2(1).

Czygan, M., Holzmann, H., & Newbold, B. (2021). Refcat: The internet archive scholar citation graph. arXiv preprint arXiv:2110.06595. Retrieved from https://arxiv.org/pdf/2110.06595

Escamilla, E., Salsabil, L., Klein, M., Wu, J., Weigle, M. C., & Nelson, M. L. (2023). It's not just github: identifying data and software sources included in publications. In International conference on theory and practice of digital libraries (pp. 195–206).

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies, 1(1), 414–427. Retrieved from https://direct.mit.edu/qss/article-pdf/1/1/414/1760913/qss a 00022.pdf

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly context not found: one in five articles suffers from reference rot. PloS one, 9 (12), e115253. Retrieved from https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0115253&type=printable

Laakso, M., Matthias, L., & Jahn, N. (2021). Open is not forever: A study of vanished open access journals. Journal of the Association for Information Science and Technology, 72(9), 1099–1112. Retrieved from https://refubium.fuberlin.de/bitstream/handle/fub188/30265/asi.24460.pdf?sequence=2&isAllowed=y

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). Open archives initiative-protocol for metadata harvesting-v. 2.0.

Ley, M. (2002). The dblp computer science bibliography: Evolution, research issues, perspectives. In International symposium on string processing and information retrieval (pp. 1–10).

markitdown. (2024). https://github.com/microsoft/markitdown

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. arXiv preprint arXiv:2402.06196.

Mohr, G., Stack, M., Rnitovic, I., Avery, D., & Kimpton, M. (2004). Introduction to heritrix. In 4th international web archiving workshop (Vol. 15, pp. 109–115).

Morrison, H. (2017). Directory of open access journals (doaj). The Charleston Advisor, 18(3), 25–28.

Peroni, S., & Shotton, D. (2020). Opencitations, an infrastructure organization for open scholarship. Quantitative Science Studies, 1(1), 428–444. Retrieved from https://arxiv.org/pdf/1906.11964

Priem, J., Piwowar, H., & Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833 . Retrieved from https://arxiv.org/pdf/2205.01833

Raasveldt, M., & Mu ̈hleisen, H. (2019). Duckdb: an embeddable analytical database. In Proceedings of the 2019 international conference on management of data (pp. 1981–1984).

Romary, L., & Lopez, P. (2015). Grobid-information extraction from scientific publications. ERCIM News, 100.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In Proceedings of the 24th international conference on world wide web (pp. 243–246). Retrieved from https://dl.acm.org/doi/pdf/10.1145/2740908.2742839

Sitemaps xml format. (2005). https://www.sitemaps.org/protocol.html

Team, D. S. (2024, 8). Docling technical report (Tech. Rep.). Retrieved from https://arxiv.org/abs/2408.09869 doi: 10.48550/arXiv.2408.09869

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). Dublin core metadata for resource discovery (Tech. Rep.).

Willinsky, J. (2005). Open journal systems: An example of open source software for journal management and publishing. Library hi tech , 23 (4), 504–519. Retrieved from https://pkp.sfu.ca/files/Library Hi Tech DRAFT.pdf