Fatcat Reference Dataset

Martin Czygan
Internet Archive
San Francisco, California, USA
martin@archive.org

Bryan Newbold
Internet Archive
San Francisco, California, USA
bnewbold@archive.org

Abstract

As part of its scholarly data efforts, the Internet Archive releases a first version of a citation graph dataset, named *refcat*, derived from scholarly publications and additional data sources. It is composed of data gathered by the fatcat cataloging project[1], related web-scale crawls targeting primary and secondary scholarly outputs, as well as metadata from the Open Library[2] project and Wikipedia[3]. This first version of the graph consists of 1,323,423,672 citations. We release this dataset under a CC0 Public Domain Dedication, accessible through an archive collection[4]. All code used in the derivation process is released under an MIT license[5].

Index terms— Citation Graph, Web Archiving

1 Introduction

The Internet Archive releases a first version of a citation graph dataset derived from a raw corpus of about 2.5B references gathered from metadata and data obtained by PDF extraction tools such as GROBID[7]. Additionally, we consider integration with metadata from Open Library and Wikipedia. The goal of this report is to describe briefly the current contents and the derivation of the dataset. We expect this dataset to be iterated upon, with changes both in content and processing.

Modern citation indexes can be traced back to the early computing age, when projects like the Science Citation Index (1955)[4] were first devised, living on in existing commercial knowledge bases today. Open alternatives were started such as the Open Citations Corpus (OCC) in 2010 - the first version of which contained 6,325,178 individual references[9]. Other notable early projects include CiteSeerX[13] and CitEc[1]. The last decade has seen the emergence of more openly available, large scale citation projects like Microsoft Academic[11] or the Initiative for Open Citations[2][10]. In 2021, according to [5] over 1B citations are publicly available, marking a tipping point for this category of data.

[1] https://fatcat.wiki
[2] https://openlibrary.org
[3] https://wikipedia.org
[4] https://archive.org/details/refcat_2021-07-28
[5] https://gitlab.com/internetarchive/cgraph

2 Related Work

There are a few large scale citation dataset available today. COCI, the "OpenCitations Index of Crossref open DOI-to-DOI citations" was first released 2018-07-29. As of its most recent release[6], on 2021-07-29, it contains 1,094,394,688 citations across 65,835,422 bibliographic resources[8].

The WikiCite[7] project, "a Wikimedia initiative to develop open citations and linked bibliographic data to serve free knowledge" continously adds citations to its database and as of 2021-06-28 tracks 253,719,394 citations across 39,994,937 publications[8].

Microsoft Academic Graph[11] is comprised of a number of entities[9] with *PaperReferences* being one relation among many others. As of 2021-06-07[10] the *PaperReferences* relation contains 1,832,226,781 rows (edges) across 123,923,466 bibliographic entities.

Numerous other projects have been or are concerned with various aspects of citation discovery and curation as part their feature set, among them Semantic Scholar[3], CiteSeerX[6] or Aminer[12].

[6] https://opencitations.net/download
[7] https://meta.wikimedia.org/wiki/WikiCite
[8] http://wikicite.org/statistics.html
[9] https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema
[10] A recent copy has been preserved at https://archive.org/details/mag-2021-06-07

Set	Count
COCI (C)	1,094,394,688
refcat-doi (R)	1,303,424,212
C ∩ R	1,007,539,966
C \ R	86,854,309
R \ C	295,884,246

Table 1: Comparison between COCI and *refcat-doi*, a subset of *refcat* where entities have a known DOI. At least 50% of the 295,884,246 references only in *refcat-doi* come from links between datasets (GBIF, DOI prefix: 10.15468).

As mentioned in [5], the number of openly available citations is not expected to shrink in the future.

3 Dataset

We release the first version of the *refcat* dataset in an format used internally for storage and to serve queries (and which we call *biblioref* or *bref* for short). The dataset includes metadata from fatcat and the Open Library Project and inbound links from the English Wikipedia.

The format contains source and target (fatcat release and work) identifiers, a few attributes from the metadata (such as year or release stage) as well as information about the match status and provanance.

The dataset currently contains 1,323,423,672 citations across 76,327,662 entities (55,123,635 unique source and 60,244,206 unique target work identifiers). The majority of matches - 1,250,523,321 - are established through identifier based matching (DOI, PMIC, PMCID, ARXIV, ISBN). 72,900,351 citations are established through fuzzy matching.

The majority of DOI based matches between *refcat* and COCI overlap, as can be seen in Table 1.

4 System Design

The constraints for the systems design are informed by the volume and the variety of the data. The capability to run the whole graph derivation on a single machine was a minor goal as well. In total, the raw inputs amount to a few TB of textual content, mostly newline delimited JSON. More importantly, while the number of data fields is low, certain schemas are very partial with hundreds of different combinations of available field values found in the raw reference data. This is most likely caused by aggregators passing on reference data coming from hundreds of sources, each of which not necessarily agreeing on a common granularity for citation data and from artifacts of machine learning based structured data extraction tools.

Fields	Percentage
CN · RN · P · T · U · V · Y	14%
DOI	14%
CN · CRN · IS · P · T · U · V · Y	5%
CN · CRN · **DOI** · U · V · Y	4%
PMID · U	4%
CN · CRN · **DOI** · T · V · Y	4%
CN · CRN · Y	4%
CN · CRN · **DOI** · V · Y	4%

Table 2: Top 8 combinations of available fields in raw reference data accounting for about 53% of the total data (CN = container name, CRN = contrib raw name, P = pages, T = title, U = unstructured, V = volume, IS = issue, Y = year, DOI = doi, PMID = pmid). Unstructured fields may contain any value. Identifiers emphasized.

Each combination of fields may require a slightly different processing path. For example, references with an Arxiv identifier can be processed differently from references with only a title. Over 50% of the raw reference data comes from a set of eight field set manifestations, as listed in Table 2.

Overall, a map-reduce style approach is followed, which allows for some uniformity in the overall processing. We extract (key, document) tuples (as TSV) from the raw JSON data and sort by key. We then group documents with the same key and apply a function on each group in order to generate our target schema or perform additional operations such as deduplication or fusion of matched and unmatched references.

The key derivation can be exact (like an identifier like DOI, PMID, etc) or based on a value normalization, like slugifying a title string. For identifier based matches we can generate the target schema directly. For fuzzy matching candidates, we pass possible match pairs through a verification procedure, which is implemented for *release entity* pairs. This procedure is a domain dependent rule based verification, able to identify different versions of a publication, preprint-published pairs and documents, which are are similar by various metrics calculated over title and authors. The fuzzy matching approach is applied on all reference documents without identifier (a title is currently required).

With a few schema conversions, fuzzy matching can be applied to Wikipedia articles and Open Library (edition) records as well. The aspect of precision and recall are represented by the two stages: we are generous in the match candidate generation phase in order to improve recall, but we are strict during verification, in order to control precision. Quality assurance for verification is implemented through a growing list of test cases of real examples from the catalog and their expected or desired match status[11].

5 Limitations and Future Work

As other dataset in this field we expect this dataset to be iterated upon.

- The fatcat catalog updates its metadata continously[12] and web crawls are conducted regularly. Current processing pipelines cover raw reference snapshot creation and derivation the graph structure, which allows to rerun processing based on updated data as it becomes available.
- Metadata extraction from PDFs depends on supervised machine learning models, which in turn depends training sets. With additional crawls and metadata available we hope to improve models used for metadata extraction, improving yield and reducing data extraction artifacts in the process.
- As of this version, a number of raw reference docs remain unmatched, which means that neither exact nor fuzzy matching can detect a link to a known entity. On the one hand, this can hint at missing metadata. However, parts of the data will contain a reference to a catalogued entity, but in a specific, dense and harder to recover form. This also include improvements to the fuzzy matching approach.

6 Acknowledgements

This work is partially supported by a grant from the *Andrew W. Mellon Foundation*. We like to thanks various teams at the Internet Archive for providing necessary infrastructure, and also data processing expertise. We are also indebted to various open source software tools and their maintainers as well as open scholarly data projects - without those this work would be much harder if possible at all.

7 Appendix A

A note on data quality: While we implement various data quality measures, real-world data, especially coming from many different sources will contain issues. Among other measures, we keep track of match reasons, especially for fuzzy matching to be able to zoom in on systematic errors more easily (see Table 3).

Count	Provenance	Status	Reason
934932865	crossref	exact	doi
151366108	fatcat-datacite	exact	doi
65345275	fatcat-pubmed	exact	pmid
48778607	fuzzy	strong	jaccardauthors
42465250	grobid	exact	doi
29197902	fatcat-pubmed	exact	doi
19996327	fatcat-crossref	exact	doi
11996694	fuzzy	strong	slugtitleauthormatch
9157498	fuzzy	strong	tokenizedauthors
3547594	grobid	exact	arxiv
2310025	fuzzy	exact	titleauthormatch
1496515	grobid	exact	pmid
680722	crossref	strong	jaccardauthors
476331	fuzzy	strong	versioneddoi
449271	grobid	exact	isbn
230645	fatcat-crossref	strong	jaccardauthors
190578	grobid	strong	jaccardauthors
156657	crossref	exact	isbn
123681	fatcat-pubmed	strong	jaccardauthors
79328	crossref	exact	arxiv
57414	crossref	strong	tokenizedauthors
53480	fuzzy	strong	pmiddoipair
52453	fuzzy	strong	dataciterelatedid
47119	grobid	strong	slugtitleauthormatch
36774	fuzzy	strong	arxivversion
35311	fuzzy	strong	customieeearxiv
33863	grobid	exact	pmcid
23504	crossref	strong	slugtitleauthormatch
22753	fatcat-crossref	strong	tokenizedauthors
17720	grobid	exact	titleauthormatch
14656	crossref	exact	titleauthormatch
14438	grobid	strong	tokenizedauthors
7682	fatcat-crossref	exact	arxiv
5972	fatcat-crossref	exact	isbn
5525	fatcat-pubmed	exact	arxiv
4290	fatcat-pubmed	strong	tokenizedauthors
2745	fatcat-pubmed	exact	isbn
2342	fatcat-pubmed	strong	slugtitleauthormatch
2273	fatcat-crossref	strong	slugtitleauthormatch
1960	fuzzy	exact	workid
1150	fatcat-crossref	exact	titleauthormatch
1041	fatcat-pubmed	exact	titleauthormatch
895	fuzzy	strong	figshareversion
317	fuzzy	strong	titleartifact
82	grobid	strong	titleartifact
33	crossref	strong	titleartifact
5	fuzzy	strong	custombsiundated
1	fuzzy	strong	custombsisubdoc
1	fatcat	exact	doi

Table 3: Table of match counts, reference provenance, match status and match reason. The match reason identifier encode a specific rule in the domain dependent verification process and are included for completeness - we do not include the details of each rule in this report.

[11]The list can be found under: `https://gitlab.com/internetarchive/cgraph/-/blob/master/skate/testdata/verify.csv`. It is helpful to keep this test suite independent of any specific programming language.

[12]A changelog can currently be followed here: `https://fatcat.wiki/changelog`

References

[1] Citations in economics. `https://citec.repec.org/`. Accessed: 2021-07-30.

[2] Initiative for open citations. `https://i4oc.org/`. Accessed: 2021-07-30.

[3] S. Fricke. Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106(1):145, 2018.

[4] E. Garfield. The evolution of the science citation index. *International microbiology*, 10(1):65, 2007.

[5] B. I. Hutchins. A tipping point for open citation data. *Quantitative Science Studies*, pages 1–5, 2021.

[6] H. Li, I. Councill, W.-C. Lee, and C. L. Giles. Citeseerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web*, pages 883–884, 2006.

[7] P. Lopez. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer, 2009.

[8] S. Peroni and D. Shotton. Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, 1(1):428–444, 2020.

[9] D. Shotton. Publishing: open citations. *Nature News*, 502(7471):295, 2013.

[10] D. Shotton. Funders should mandate open citations. *Nature*, 553(7686):129–130, 2018.

[11] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246, 2015.

[12] J. Tang. Aminer: Toward understanding big scholar data. In *Proceedings of the ninth ACM international conference on web search and data mining*, pages 467–467, 2016.

[13] J. Wu, K. Kim, and C. L. Giles. Citeseerx: 20 years of service to scholarly big data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, pages 1–4, 2019.