Refcat: The Internet Archive Scholar Reference Graph

2023-02-17 / Martin Czygan, Internet Archive / Kompetenzzentrum Bibliometrie / Berlin / https://bibliometrie.info

Martin Czygan

- /about: Open data engineer, Software developer
- /affiliations: Internet Archive, Leipzig University Library
- /etc: Open Source, Writing, Consulting, Teaching

Work presented by me, but collaborative effort at the Internet Archive with Bryan Newbold, Helge Holzmann, Jefferson Bailey (PI) and others.

Background / Open is not forever

- Scholarly communications artifacts as critical archival subjects
- Since 2017, two projects at the Internet Archive (funded partially by the Mellon Foundation)

Open is not forever: A study of vanished open access journals (10.1002/asi.24460, 2021)

- Follow up issues, e.g. **citation integrity** (papers and beyond)

Background / Implementation

Implementation

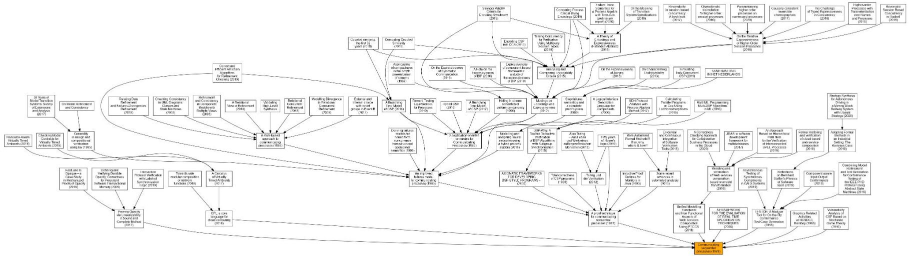
- collect and catalogue metadata (fatcat)
- archive full-text
- archive additional artifacts, like datasets
- access (https://fatcat.wiki, https://scholar.archive.org)

A set of harvesters, indexers and archiving workflows for **continuous updates** (https://fatcat.wiki/changelog) as well es targeted large-scale web-crawls.

- open source at https://github.com/internetarchive/fatcat, https://github.com/internetarchive/fatcat-scholar

Background / Results

- Millions (and TBs) of papers, datasets preserved and catalogued (ongoing)
- **Internet Archive Scholar** https://scholar.archive.org/ access site and search over 25M full-texts and 100M+ metadata records (since 03/2021)
- **Citation Graph** (refcat) as data derivation (v1 in 10/2021, v3 in progress)



Refcat / Overview

- a **open citation dataset** derived from archived metadata and full-text analysis – using both **id-based** and **fuzzy matching** techniques
- download latest version via: https://archive.org/details/ia_biblio_metadata?query=refcat
- open source: https://gitlab.com/internetarchive/refcat/
- documented (v1) in preprint: https://arxiv.org/abs/2110.06595 (2021)
- blog post: https://blog.archive.org/2021/10/19/internet-archive-releases-refcat-the-ia-scholar-index-of-over-1-3-billion-scholarly-citations/

Refcat / Details

- currently (v2): 1,462,333,688 (doi-doi)
- most edges found via id-based matching
- about 5% of the edges come from fuzzy matching
- we include **outbound links** to Open Library (3M+ books; which books are referenced in papers?)
- we include **inbound links** from (en) Wikipedia (6M+ articles)

See also: A tipping point for open citation data (10.1162/qss_c_00138, 2021)

Refcat / Paper Outlinks

- we analyzed links (URLs) found in papers
- in a sample (from 10/2021) of 364415 URLs found in papers **and** preserved at the Internet Archive, about 16% were not accessible on the live-web anymore
- preservation of scholarly communications artifacts critical for **citation integrity**

Refcat / Process

The whole project used a mixed top-down (open metadata) and bottom-up (archived material) approach.

- essential data aggregators: crossref, datacite, doaj, ...
- specific data providers: arxiv, dblp, ...
- IR and journal harvesting: 70k+ endpoints (metha)
- IA collections

A variety of datasets collected in the process: https://archive.org/details/ia_biblio_metadata

Refcat / Process

- find reference information in metadata
- use GROBID / wapiti to process raw PDF files to extract references
- combine all reference data into a single file, one line per reference (about 2.5B lines)
- get Open Library and Wikipedia snapshots
- analyze input data and synthesize citation dataset in a processing pipeline

Refcat / Derivation

- use identifiers (doi, arxiv id, pubmed id, ...) - if no ID found to match, use **fuzzy matching** over title and various other fields (candidate generation and verification)
- large scale processing of billions of records (on a single machine, w/ Go)
- released as a single file snapshot: https://archive.org/details/ia_biblio_metadata?query=refcat
- Latest snapshot: v2 (v3 in progress)

Refcat / Observations

- Data is very messy
  - inhibits faster progress
  - data quality requirements (do not want bad links)
  - obscure ways to express a reference
  - manual verification process
  - balance between performance and matching techniques (scalable, yet lightweight solutions desired)
- Lots of little improvements possible
  - matching more Internet Archive holdings to catalog entities
  - experiment with new match-key algorithms
  - uncover and express reference patterns (“funnel approach”)

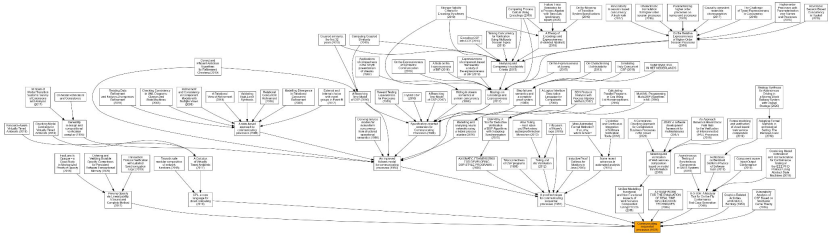
Outlook

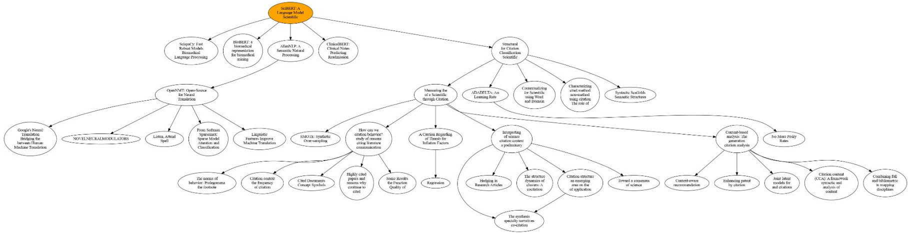
Ongoing and future tasks:

- Continuous metadata acquisition and harvesting of the web, referenced URLs and related content
- Citation graph derivations with data and processing updates
- Citations graph diversity: webpages, wikipedia articles, books, datasets and other referenceable entities
- Refinements of the matching process
- Detailed comparison of various open citation datasets (like OpenAlex, OpenCitations)

Outlook / Applications

- Prototypical reference graph and library catalog merging at SLUB Dresden: Project LABE (talk)
- CLI utilities to render graphs (e.g. with graphviz)





SciBERT: A Language Model Scientific
ScispaCy: Fast Robust Models Biomedical Language Processing
BioBERT: a biomedical representation for biomedical mining
AllenNLP: A Semantic Natural Processing
ClinicalBERT: Clinical Notes Predicting Readmission
Structural for Citation Classification Scientific
OpenNMT: Open-Source for Neural Translation
Measuring the of a Scientific through Citation
ADADELTA: An Learning Rate
Contextualizing for Scientific using Word and Domain
Characterizing cited method non-method using citation The role of
Syntactic Scaffolds Semantic Structures
Google's Neural Translation Bridging the between Human Machine Translation
NOVELNEURALMODULATORS
Listen, Attend Spell
From Softmax Sparsemax: Sparse Model Attention and Classification
Linguistic Features Improve Machine Translation
SMOTE: Synthetic Over-sampling
How can we citation behavior? study of reasons citing literature communication
A Caution Regarding of Thumb for Inflation Factors
Interpreting of science citation context a preliminary
Content-based analysis: The generation citation analysis
No More Pesky Rates
The norms of behavior: Prolegomena the footnote
Citation context the frequency of citation
Cited Documents Concept Symbols
Highly cited papers and reasons why continue to cited
Some Results the Function Quality of
Regression
Hedging in Research Articles
The structure dynamics of clusters: A cocitation
Citation structure an emerging area on the of application
Toward a consensus of science
Context-aware recommendation
Enhancing patent by citation
Joint latent models for and citations
Citation content (CCA): A framework syntactic and analysis of content
Combining full and bibliometric in mapping disciplines
The synthesis specialty narratives co-citation

Thank you, contact us!

- we are committed to open source and the open data ecosystem
- other presentations:
  - https://www.youtube.com/watch?v=PARqfbYIdXQ (Perpetual Access Machines: Archiving Web-Published Scholarship at Scale - Jefferson Bailey)
- Martin Czygan, Open data engineer, martin@archive.org