From Open Access to Perpetual Access: Archiving Web-Published Scholarship

Maria Praetzellis
Program Manager, Web Archiving & Data Services | Internet Archive
IIPC WAC 2019 | maria@archive.org
Outline

1. Archiving (Digital) Scholarship
2. Conceptual Approaches
3. Technical Approaches
4. Fatcat Beta Walkthrough
5. Fat Machine Learning Cat
Outline

1. Archiving (Digital) Scholarship
Archiving Digital Scholarship One-Liner

Build a complete, use-oriented, highly-available archive and knowledge graph of every publicly-accessible scholarly output + descriptive metadata and full-text, linked with versions and secondary outputs (data/blogs/etc) with a priority on long-tail, at-risk publications -- all accessible via API-first editable, distributed catalog that includes links to files in the web archive.
Goals/Concepts of this Work

- Apply automation & scale of web harvesting to archiving specific content (scholarly works)
- Extract and add metadata to improve discovery of those resources in web archives
- Apply above to past web archives
- Use machine learning to improve processes
- Provide API-first access to this corpus
- Provide non-profit, open infrastructure for perpetual access to knowledge
Some Numbers

1. There are ~150-200M scholarly articles
   a. How can we get all that are on the web
   b. Once archived, how can we make all discoverable w/o knowing (wayback) URL

2. There are ~600M PDFs in Wayback Machine
   a. How can we know which are scholarship
   b. Once known, how can we make those discoverable w/o knowing (wayback) URL
Outline

1. Archiving (Digital) Scholarship
2. Conceptual Approaches
Conceptual Approaches 1

1. Identifier & metadata services (DOIs, ISSNs, etc) contain URLs of scholarly works
   a. We will archive the metadata and the URLs

2. Web-scale harvesting is cheap in time/resources to archive ten/hundred millions of scholarly works
   a. Automate for “scrape-to-crawl-to-find” process

3. Many efforts are aggregating scholarship but not for perpetual access and not the long-tail stuff
   a. Advance work via partnerships, manifests sharing, system/service integrations
Conceptual Approaches 2

● Top-down:
  ○ Use lists/IDs/MD/etc to target harvesting and associate scholarship with metadata
  ○ Extract metadata from archived works

● Bottom-up:
  ○ ML/algorithms to identify scholarly works already in the archive, assess quality of preservation of a web-only publication
  ○ ML/algorithms to identify, archive, and associate “secondary” works (data, blog, etc)
Outline

1. Archiving (Digital) Scholarship
2. Conceptual Approaches
3. Technical Approaches

IIPC WAC 2019
Sources

- Manifests: Unpaywall, CORE (UK), ISSN, Semantic Scholar, DOAJ, MS Academic, CiteSeerX, Meta, other
- Metadata: DOIs (CrossRef), ISSNs, ORCIDs, DataCite, Wikidata, PubMed, etc
- Other: SHERPA/RoMEO (license); Keeper’s Registry (preservation)
Partnerships

Find and understand peer-reviewed research papers

Try: "vaccines and autism," "How safe is bicycling," or "chocolate"
APIs, Reporting, Bulk Access
A Large-Scale Analysis of Impact Factor Biased Journal Self-Citations

by Caspar Chorus, Ludo Waltman

Date (published): 2016-08-25
PubMed: 27560867
PubMed Central: PMC4999059
Wikidata Entity: Q36113005

This journal article is a release (version) of the work t2q7ttx4tref7boytqswxyefay. There may be other releases (e.g. pre-prints, formal publications, etc) linked to the same work.

Published in PLoS ONE by Public Library of Science (PLoS)

Extra Metadata (raw JSON)
crossref: <truncated, see full JSON>

Abstracts
No known abstracts.

All Contributors

<table>
<thead>
<tr>
<th>Attribution Order</th>
<th>Name</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Caspar Chorus</td>
<td>author</td>
</tr>
<tr>
<td>2</td>
<td>Ludo Waltman</td>
<td>author</td>
</tr>
<tr>
<td></td>
<td>Wolfgang Glanzel</td>
<td>editor</td>
</tr>
</tbody>
</table>

Known Files and URLs

<table>
<thead>
<tr>
<th>SHA-1</th>
<th>Size (bytes)</th>
<th>File Type</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>21471</td>
<td></td>
<td>application/pdf</td>
<td>repository.tudelft.nl (web)</td>
</tr>
<tr>
<td>53545</td>
<td></td>
<td>application/pdf</td>
<td>journals.plos.org (web)</td>
</tr>
</tbody>
</table>
Outline

1. Archiving (Digital) Scholarship
2. Applying Web Archiving Methods
3. Conceptual Approaches
4. Technical Approaches
5. Fatcat Beta Walkthrough
Fatcat! (Big Catalog)

- Editable catalog tracking the (archival) location, metadata, and status of research objects to ensure perpetual access
- Built by matching crawled web content (both historical and ongoing) against metadata
- Now at ~150M metadata records, ~18M known full text works, ~70M likely total works, ~700M citations
Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

by Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, Richard Tobin

- Published in PLoS ONE by Public Library of Science (PLoS)
- All Contributors (8)

Extra Metadata (raw JSON)

crossref.type: journal-article

crossref.license: ['start': '2014-12-26T00:00:00Z', 'URL': 'http://cr...']

Known Files and URLs

application/pdf 1.8 MB

web.archive.org (webarchive)
web.archive.org (webarchive)
www.plosone.org (web)
journals.plos.org (publisher)
web.archive.org (webarchive)
+ 5 more URLs

References

This release citing other releases

7. Wavelab and reproducible research Wavelets and Statistics.199555 (DOI: 10.1007/978-1-4612-2544-7_5)

https://fatcat.wiki/
Wayback(!) and live web URLs + mime, size, checksum

Extracted citations (interlinked to other fatcat records and wayback URLs for web references)

https://fatcat.wiki/
The API, which has additional metadata not in the user interface.
Outline

1. Archiving (Digital) Scholarship
2. Applying Web Archiving Methods
3. Conceptual Approaches
4. Technical Approaches
5. Fatcat Beta Walkthrough
6. Fat Machine Learning Cat
FatMLCat Goals

Build classifiers that:

● Identify scholarly articles in web archives
● Identify whether online scholarly publications are being well archived (improve if not)
● Identify unknown online scholarly publications not being archived (and archive them)
● Apply fatcat process to these resources for improved discovery and distribution
FatMLCat Specifics

● Is this PDF/HTML a scholarly article?
  ○ Signals: host name or URL string; doc format or layout; analyze & compare metadata, login page and “partial copy” detectors

● Is this online scholarly publication “well archived”?
  ○ Signals: estimate correct capture frequency, size, number; model content type, flags for variance

● How can we find and archive online long-tail scholarly sites we don’t know about
  ○ Signals: link graph, citation graph
FatMLCat Outcomes

● Technicals: Using Spark MLlib, scikit-learn, with most code in Scala or Python
● Improvement of existing open source tools in the fatcat/fatMLCat workflow (GROBID, etc)
● All training sets, classifiers, and code will be released open source in early 2020
● Will also release cost models on the costs (per TB) to run similar jobs, local or cloud
FatMLCat to the Future

- Run classifiers on multiple ccTLD full domain crawls
- Run classifiers on multiple university *.edu crawls in Archive-It
- [Thanks partners! Others welcome!]
- Services for IDing and MDing scholarship in domain/host crawls
- Services to deliver these subsets or relevant off-domain/host subsets to partners for local use/preservation
- Computational research services
Further Thoughts & Light Reading

Thoughts:
- Leverage WA methods for all preservation/access stuff
- Better knowledge/discovery of what’s in web archives
- Delivery of relevant subsets into web archives / IRs

Readings:
- “Andrew W. Mellon Foundation Awards Grant to the Internet Archive for Long Tail Journal Preservation”
  - https://blog.archive.org/ (search “mellon”)
- “Personal Pods and Fatcat,” DSHR blog
- Fatcat announcements upcoming on IA blog
THANKS!
CONTACT IF INTERESTED!

Jefferson Bailey, jefferson@archive.org
Director, Web Archiving & Data Services

Maria Praetzellis, maria@archive.org
Program Manager, Web Archiving & Data Services

Credits: Bryan Newbold (FatCat Open Data Engineer)
Volunteers: David Rosenthal, Vicky Reich
Partial Funding: Mellon Foundation

Internet Archive
https://archive.org

Archive-It
https://archive-it.org

https://fatcat.wiki

IIPC WAC 2019