Perpetual Access Machines: Archiving Web-published Scholarship at Scale



Jefferson Bailey, Director, Web Archiving & Data Services, Internet Archive Bryan Newbold, Open Data Engineer, Web Archiving & Data Services, Internet Archive FORCE11 2019 | EDI | [jefferson], [bnewbold] @archive.org



Non-Profit Digital Library Founded in 1996 HQ IN SF (come visit!)

Universal Access to All Knowledge









Collection / Preservation Stats

Total Archiving ~8PB Total collected/year **~30TB** Total collected/day ~50% Web data Web & Data Services "2.x+PB Per year via paid services 600 & \$6M Orgs & Rev **~3.xTB** Per year per org <u>average</u>



Outline

1. At Risk Scholarship 2. Goals & Methods 3. Technical Approaches 4. What It Looks Like 5. Partnerships, Roadmap, Services



Perpetual Access Archiving Challenges

• Print >>> Digital = custodial challenges Traditional curation doesn't scale Traditional deposit/services obsolete **First-world preservation non-problems** Long Tail: not English, STEM, or US/EU 94 problems... for web/born-digital OA







Perpetual Access Access Challenges

F11 2019

Wayback Machine = known URLs But has ~600M PDFs (~5% scholarly) • Web archives not (yet) in most search Scholarly outputs all over the dang web • No QA/QC mechanisms or edit tools (Prior) no targeted, integrated harvesting



Outline

At Risk Scholarship
 Goals & Methods

Technical Approaches
 What It Looks Like
 Partnerships, Roadmap, Services





IA open infrastructure for public good/access Instead of specialized archiving services... Add "open scholarship" automation to the existing automation/scale of web harvesting Instead of specialized curation and ingest... Develop tools to identify scholarly objects in existing harvests/archives, ID/correct incompleteness, augment with metadata F11 2019

ARCHIVE-

Methods / Approaches

Top-Down Approach Harvest/archive PIDs, registries, metadata, manifests, etc to find web-published scholarly outputs to archive, extract/augment with metadata, etc



Methods / Approaches Bottom-Up Approach Use machine learning tools to identify scholarly objects in TB/PB scale web archives, assess completeness, and match with versions, metadata, etc





Methods / Approaches

Forever Approach **Partnerships, services Open APIs, code, infrastructure** Add to discovery services Add to data services **Re-distribution & bulk access**





Outline

1. At Risk Scholarship 2. Goals & Methods **3.** Technical Approaches 4. What It Looks Like 5. Partnerships, Roadmap, Services



Tech Constraints

 Bots and humans working together Augmentation **Open to public contributions** o don't be a bottleneck **Build little new IA infrastructure** efficient, risk reduction, perpetual-ish





Ingest: Crawl Existing Indices





arXiv.org





Cite

Usual suspect (registrars, indices)

- Transform metadata into CSL-like schema, import
- Bulk URL seedlist crawling, import "hits"
- Work with peer aggregators hosting fulltext • OAI-PMH / IRs











Ingest: Index Web Archives

Petabytes of unsorted web resources

 HTML, PDF, Datasets, everything

 Filter down to likely research pubs
 Try matching against existing catalog
 Else, create new catalog records







pdf classify



PDF Tooling

Application of Bic cycle assessment to do investigating on production of used and forth trajec (dotal) producers:

grobid

<?xml version="1.0"?>
<quiz>
<quida seq="1">
<question>
Who was the forty-second
president of the U.S.A.?
</question>
<answer>
William Jefferson Clinton
</answer>
</qanda>
<!-- Note: We need to add</pre>

more questions later.--> </quiz>

biblio-glutton

PG3557 .R5355 Grisham, John F57 1991

> The firm / John Grisham. 1st. ed. New York : Doubleday, c1991. 421p. ; 24 cn.

Government investigators--Fiction.
 Organized crime--Fiction.





Outline

1. At Risk Scholarship 2. Goals & Methods 3. Technical Approaches 4. What It Looks Like 5. Partnerships, Roadmap, Services



Search Papers...

https://fatcat.wiki/

Perpetual Access to Millions of Open Research Publications From Around The World

by title, authors, identifiers...

Search

140.085

Journals

96,947,165 Papers 18,117,429 Fulltext

Fatcat is a versioned, user-editable catalog of research publications including journal articles, conference proceedings, and datasets

Features include archival file-level metadata (verified digests and long-term copies), an **open**, **documented API**, and work/release indexing (eg, distinguishing between and linking pre-prints, manuscripts, and version-of-record). Read more...

This service is hosted at **The Internet Archive**, a US non-profit dedicated to providing Universal Access to All Knowledge. Donations welcome!

Development funding comes from The Andrew Mellon Foundation to improve preservation and access to "long-tail" open access works on the public web which might otherwise be lost.



Basic Reader Access



ARCHIVE

ARCHIVE-I1

Hypertext research has primarily focused on a single document or set of related documents converted to

F11 2019

Wayback Replay

The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles

release hsmo6p4smrganpb3fndaj2lon4

by Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, Stefanie Haustein

Overview Authors (9)

References (52)

Edit History

Abstract



Mirror Locations



<jats:p>Despite growing interest in Open Access (OA) to scholarly literature, there is an unmet need for large-scale, up-to-date, and reproducible studies assessing the prevalence and characteristics of OA. We address this need using oaDOI, an open online service that determines OA status for 67 million articles. We use three samples, each of 100,000 articles, to investigate OA in three populations: (1) all journal articles assigned a Crossref DOI, (2) recent journal articles indexed in Web of Science, and (3) articles viewed by users of Unpaywall, an open-source browser extension that lets users find OA articles using oaDOI. We estimate that at least 28% of the scholarly literature is OA (19M in total) and that this proportion is growing, driven particularly by growth in Gold and Hybrid. The most recent year analyzed (2015) also has the highest percentage of OA (45%). Because of this growth, and the fact that readers disproportionately access newer articles, we find that Unpaywall users encounter OA quite frequently: 47% of articles they view are OA. Notably, the most common mechanism for OA is not Gold, Green, or Hybrid OA, but rather an under-discussed category we dub Bronze: articles made free-toread on the publisher website, without an explicit Open license. We also examine the citation impact of OA articles, corroborating the so-called open-access citation advantage: accounting for age and discipline, OA articles receive 18% more citations than average, an effect driven primarily by Green and Hybrid OA. We encourage further research using the free oaDOI service, as a way to inform OA policy and practice.</jats:p>

Metadata

In application/xml+iats format

Published in PeerJ by PeerJ

Known Files and URLs

application/pdf 2.4 MB sha1:bca1531b0562c6d72e0c... web.archive.org (webarchive) peeri.com (web)

Read Full Text

Type article-journal Stage published Date 2018-02-13

DOI 10.7717/peerj.4375 PubMed 29456894 PMC PMC5815332 Wikidata 049873702

Container Metadata Open Access Publication ✓ In DOAJ In ISSN ROAD @ ISSN-L: 2167-8359

Work Entity grouping other versions (eg, pre-print) and variants of this release

Lookup Links

Fatcat Bits State is "active". Revision: de98deaf-4720-430a-8a97-1ff244cb602f As JSON object via API

View History Edit Metadata

Publication Status

PIDs Galore

Journal Metadata

F11 2019

Alt. Versions

API Helpers Wiki Sauce

Editing Metadata



⊢ A R C H I V E

ARCHIVE-IT

-



description url; title formatting

Creator Edits (0)

File Edits (0)

File Set Edits (0)

Web Capture Edits (0)

Comments and Annotations

Length Strategy at 2019-10-11 01:34:12	
This is an example comment on an editgroup. Hooray revi	ew!
Add Comment	
Markdown is allowed	😰 Submit

Collaborative Review

	Recent Ch	nanges	
	Limited to the most recent entries.		
	Changelog Index	Editgroup	Description
	2019-10-11 01:31:57	<pre> fr crossref-bot 75a64wbih5cy7oq2jeihxfelr4 </pre>	Automated import of Crossref DOI metadata, harvested from REST API
	#2976547 2019-10-11 01:31:56	🙀 crossref-bot gmfiq5lwufeftKkpajmdni2c4u	Automated import of Crossref DOI metadata, harvested from REST API
	#2976546 2019-10-11 01:31:55	<pre></pre>	Automated import of Crossref DOI metadata, harvested from REST API
A E	#2976545 2019-10-11 01:31:54	ℜ crossref-bot yixqnizhyfh5rjbbbe5vhtqbaa	Automated import of Crossref DOI metadata, harvested from REST API
	#2976544 2019-10-11 01:31:52		Automated import of Crossref DOI metadata, harvested from REST API
	#2976543 2019-10-11 01:31:51	<pre>g crossref-bot 3ivvugcdqnbu3mrcgwf6umviuy</pre>	Automated import of Crossref DOI metadata, harvested from REST API
	#2976542 2019-10-11 01:31:50	<pre> crossref-bot qe2vb7qevnglfc5r2q6q5e273a </pre>	Automated import of Crossref DOI metadata, harvested from REST API
	#2976541 2019-10-11 01:31:49	ℜ crossref-bot ap?amyk3cncu3fwp4iyiu5xgrm	Automated import of Crossref DOI metadata, harvested from REST API
Contraction of the local division of the loc	#2976540 2019-10-11 01:31:48	<pre></pre>	Automated import of Crossref DOI metadata, harvested from REST API
Constanting of the	#2976539 2019-10-11	♣ crossref-bot	Automated import of Crossref DOI

Changelog

F11 2019

Coverage Dashboards



ARCHIVE-IT

#DeveloperThings



PUT /editgroup/{editgroup_id} Request samples Payload

Velcome Editing Quickstart 1.2. Data Model 1.3. Editing Workflow 1.4. Sources of Metadata 1.6. Roadman Cataloging Style Guide 2.1. All Entitie 2.2. Containe 2.3. Creator 2.4 File 2.5. Fileset 2.6 Web Capture 27 Pelease 2.8. Work Dublic ADI 3.1. Bulk Exports

5.1. Code of Conduc 5.2 Privary

ibliography About This Guide

11 Goals and Related Projects 1.5. Implementation and Infrastructure

3.2. Cookboo

Cookbook

These quickstart examples gloss over a lot of details in the API. The canonical API documentation (generated from the OpenAPI specification) is available at https://api.fatcat.wiki/redoc.

The first two simple cookbook examples here include full headers. Later examples only show python client library code snippets.

Lookup Fulltext URLs by DOI

Often you have a DOI or other paper identifier and want to find open copies of the paper to read. In fatcat terms, you want to lookup a release by external identifier, then sort through any associated file entities to find the best files and LIRI's to download. Note that the Linnawall API is custom designed for this task and you should look in to using that instead.

This is read-only task and requires no authentication. The simple summary is to:

- 1. GET the release lookup endpoint with the external identifier as a query parameter. Also set the hide parameter to elide unused fields, and the expand parameter to files to include related files in a single request.
- 2. If you get a hit (HTTP 200), sort through the files field (an array) and for each file the urls field (also an array) to select the hest LIRL(s)

The URL to use would look like https://api.fatcat.wiki/v0/release/lookup?doi=10.1088/0264 9381/19/7/380&expand=files&hide=abstracts.refs in a browser. The query parameters should be URL encoded (eg. the DOI / characters replaced with with 120), but almost all HTTP tools and

The raw HTTP request would look like:

libraries will do this automatically

GET /v0/release/lookup?doi=10.1088%2F0264-9381%2F19%2F7%2F380&expand=files&hide=abstr Accept: */* Accept-Encoding: gzip, deflate

F11 2019

Connection: keep-alive Host: api, fatcat, wiki User-Agent: HTTPie/0.9.8

API Specification and Docs https://api.fatcat.wiki

"The Guide" https://guide.fatcat.wiki



Also: dumps, feed, code, client libraries, sandbox, etc

Outline

1. At Risk Scholarship 2. Goals & Methods 3. Technical Approaches 4. What It Looks Like 5. Partnerships, Roadmap, Services



PARTNER

dat://

ARCHIVE

ARCHIVE-I

Center for Open Science - open data archiving PID+, preservation, & discovery services Unpaywall/OR: data/access sharing **CDL/Dat: datasets in dweb** Partial funding: Mellon, IMLS Semantic Scholar our research **OPEN SCIENCE**

ISSN



THE ANDREW W. MELLON FOUNDATION

Museum and Library SERVICES

F11 2019



JOURNALS

Current Work & Roadmap

Beyond PDF: html, data, code, etc
Secondaries & platforms
"Save Paper Now"
Full-text search for everything
Citation integrity & indexing
Bulk Data Sharing

ARCHIVE

ARCHIVE-IT

Entities "Papers" Total 98,142,998 Fulltext on web 18,117,430 "Gold" Open Access 9,700,221 In a Keepers/KBART archive 59,742,897 On web, not in Keepers 8,203,985 Releases Total 121,015,771 References (raw, unlinked) 780,160,922 Containers Total 140,931

F11 2019

https://fatcat.wiki/stats



Service Developments

• Build on existing service relationships with 600+ research & knowledge orgs • Embed in existing services, Archive-It **Content Deposit Services Data/Digital Preservation Services Computational Research Data Services PIDtegrations** • We'll do any pilots, collaborations, R&D!





THANKS! CONTACT US!

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

Bryan Newbold, Open Data Engineer bnewbold@archive.org

Special Thanks!! Volunteers: David Rosenthal, Vicky Reich, Ellen Spertus Funders: Mellon Foundation, IMLS



Internet Archive, <u>https://archive.org</u> Fatcat beta, <u>https://fatcat.wiki/</u>



THANK YOU