

Perpetual Access Machines: Archiving Web-published Scholarship at Scale

[illegible]

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



The Internet Archive

Non-Profit Digital Library

Founded in 1996

HQ IN SF (come visit!)

Universal Access to All Knowledge

Total Digital Items

200,000 Software Titles

3,000,000 Moving Images

3,500,000 Books

4,500,000 Audio Recordings

20,000,000 Texts*

1,000,000 Users/day

800+ Institutional Partners

800,000,000,000+ URLs

50 PB (unique)

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 average

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**

Access

Perpetual Access Access Challenges

- **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

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

goal goal

- 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**

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

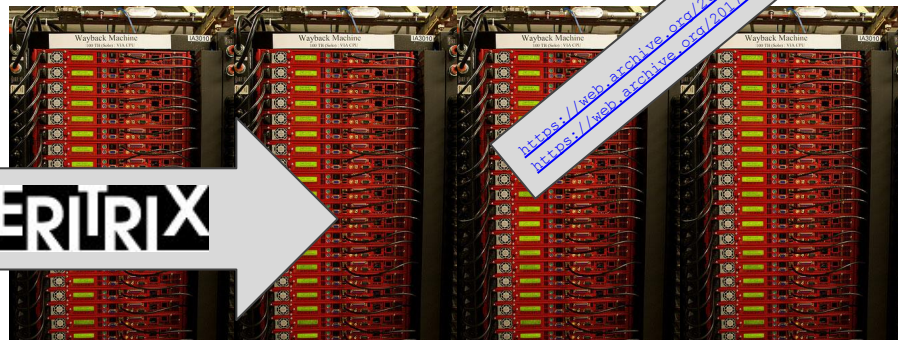
- **Bots and humans working together**
 - Augmentation
- **Open to public contributions**
 - don't be a bottleneck
- **Build little new IA infrastructure**
 - efficient, risk reduction, perpetual-ish



"sandcrawler"



WayBackMachine



"fatcat"



<https://web.archive.org/2018010203/http://plos.org/>
<https://web.archive.org/2018010203/http://eife.org/>



<https://fatcat.wiki/>

<https://github.com/internetarchive/fatcat>



Ingest: Crawl Existing Indices



- 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





Bulk Bibliographic Metadata

Internet Archive Web Group

This collection contains both external ("upstream") metadata dumps and Internet Archive generated databases and reports on our holdings of papers, books, and other documents.

- Share
- Favorite
- RSS
- Edit
- History
- Play All

ABOUT

COLLECTION

82 RESULTS

Search this Collection

- Metadata
- Text contents

PART OF

The Internet Archive

Media Type

- data 81
- texts 1

Year

- 2019 17
- 2018 22
- 2017 11
- 2016 3
- 2015 1
- 2014 1

More

Topics & Subjects



SORT BY VIEWS TITLE DATE

https://archive.org/details/ia_biblio_metadata

Crossref
Crossref DOI Dump (2018-01)
459 1 0

Crossref
Crossref DOI Dump (2018-09)
259 0 0

Microsoft Academic Search
Microsoft Academic Graph (2016-02-05 snapshot)
by Microsoft Academic Search
233 0 0

oadoi
oadoi DOI/URL Dataset
226 0 0

Sci-Hub DOI List
by Sci-Hub
225 0 0

INTERNET ARCHIVE
Internet Archive Paper Manifest (2017-09-19)
151 0 0

INTERNET ARCHIVE
Internet Archive Paper Manifest (2018-01-25)
136 0 0

CiteSeerX
CiteSeerX Database Dump (2017-03-31)
by CiteSeerX Group at PSU
128 0 0

Aminer
Open Academic Graph (aminer.org)
by aminer.org
121 0 0

ORCID
ORCID Public Data File (2017)
by ORCID, Inc.
84 0 0

ROAD
ROAD/ISSN Directory (2018)
by ROAD: Directory of Open Access Scholarly Resources
84 0 0

DOAJ
DOAJ Journal and Article Metadata (2018)
by Directory of Open Access Journals
71 0 0

CORE
CORE Open Access Paper Metadata (2017-11-)
68 0 0

INTERNET ARCHIVE
Internet Archive Paper Manifest (2017-10-06)
65 0 0

Journal ISSN Metadata Exploration (2018-04-05)
by Internet Archive Web Group
Pie chart showing HTTP Status (pre-redirect): 48.0% (200), 47.0% (3xx), 5% (404), 5% (5xx).

Ingest: Index Web Archives

1. **Petabytes of unsorted web resources**
 - HTML, PDF, Datasets, everything
2. **Filter down to likely research pubs**
3. **Try matching against existing catalog**
4. **Else, create new catalog records**

PDF Tooling



pdf_classify

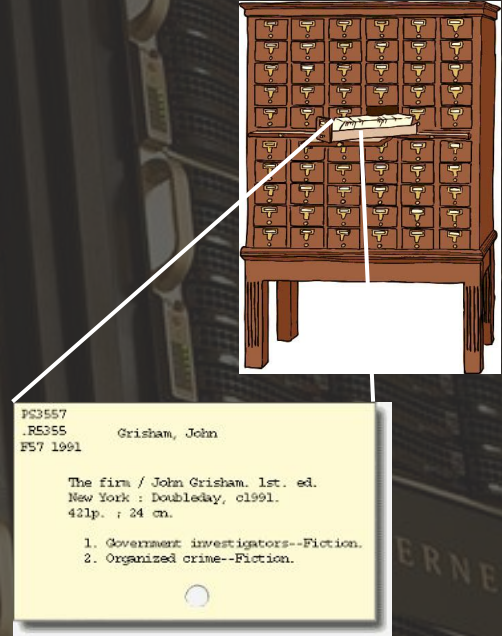


grobid

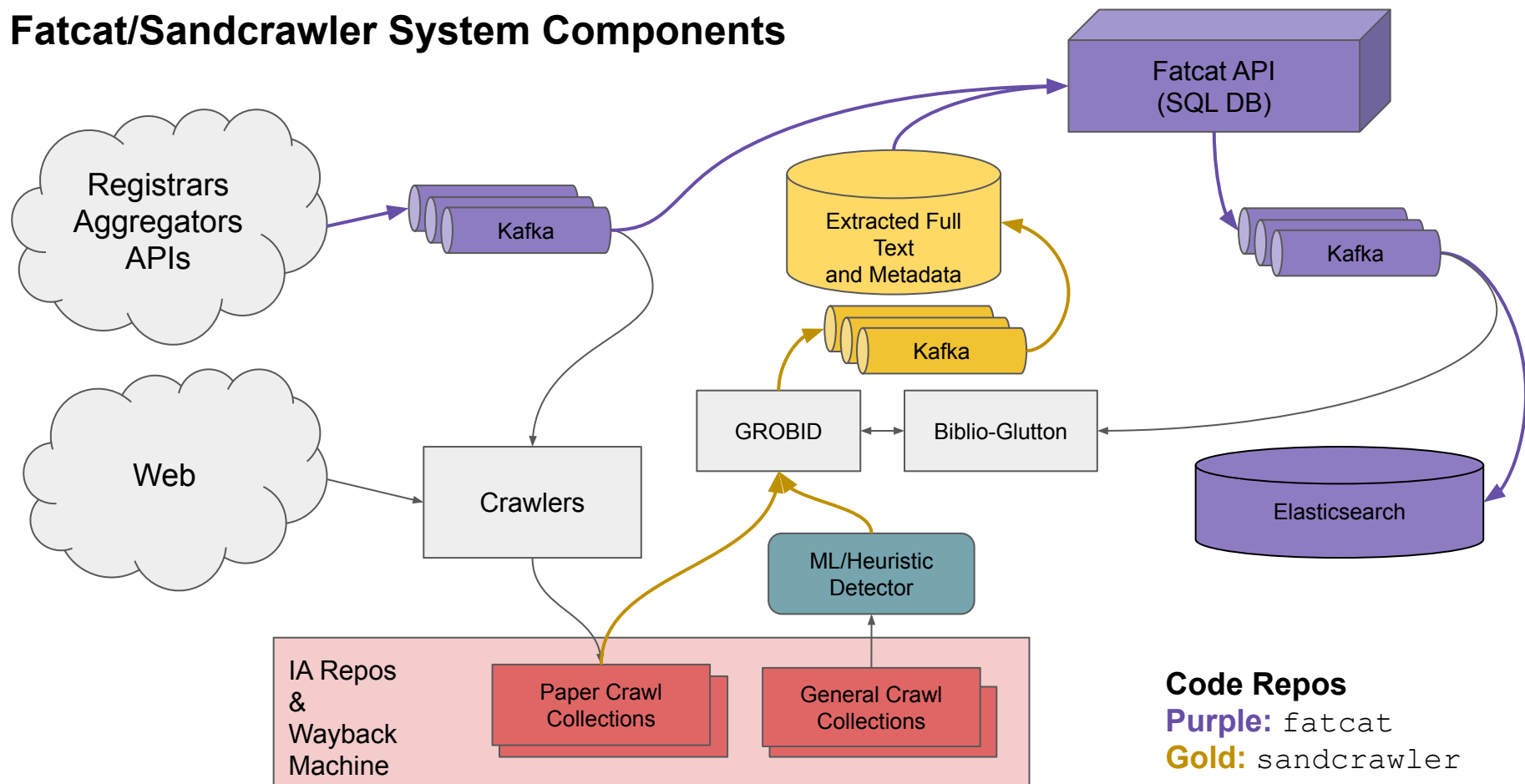
```
<?xml version="1.0"?>
<quiz>
  <qanda seq="1">
    <question>
      Who was the forty-second
      president of the U.S.A.?
    </question>
    <answer>
      William Jefferson Clinton
    </answer>
  </qanda>
  <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

biblio-glutton



Fatcat/Sandcrawler System Components



Outline

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

<https://fatcat.wiki/>

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

by title, authors, identifiers...

Search

96,947,165
Papers

18,117,429
Fulltext

140,085
Journals



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

fatcat! About Guide Changelog Search Papers... Login/Signup

Search all Releases

ellen spertus

Can also lookup by identifier or search for containers (eg, journals). ☐ Fulltext Available Only

Showing top 23 out of 23 results for: ellen spertus

ParaSite: mining structural information on the Web
Ellen Spertus
1997 Computer networks and ISDN systems
doi:10.1016/S0169-7552(97)00033-0

Dataflow Computation for the J-Machine
Ellen Spertus
1990
doi:10.21236/ada228612

Gender benders
Ellen Spertus
2002 ACM SIGCSE Bulletin
doi:10.1145/543812.543848

Leveraging an alternative source of computer scientists
Sheila Humphreys, Ellen Spertus
2002 ACM SIGCSE Bulletin
doi:10.1145/543812.543830

Squeal: a structured query language for the Web
Ellen Spertus, Lynn Andrea Stein
2000 Computer Networks
doi:10.1016/S1389-1286(00)00074-8

Evaluating the locality benefits of active messages
Ellen Spertus, William J. Dally
1995 Proceedings of the fifth ACM SIGPLAN symposium

Search

INTERNET ARCHIVE waybackmachine <http://people.mills.edu/spertus/Papers/parasite97.pdf> Go SEP 18 APR 18 JUL 18 10 captures 18 Dec 2006 - 18 Aug 2017 2006 2007 2008 About this capture

ParaSite: Mining Structural Information on the Web 1 / 13

http://www.mills.edu/ACAD_INFO/MCS/SPERTUS/Parasite/parasite.html

Appearing in *The Sixth International World Wide Web Conference*, April 1997.

ParaSite: Mining Structural Information on the Web

Ellen Spertus
MIT Artificial Intelligence Lab and University of Washington Dept. of CSE
University of Washington
Box 352350
Seattle, WA 98195-2350
ellens@ai.mit.edu

Abstract

Web information retrieval tools typically make use of only the text on pages, ignoring valuable information implicitly contained in links. At the other extreme, viewing the Web as a traditional hypertext system would also be a mistake, because heterogeneity, cross-domain links, and the dynamic nature of the Web mean that many assumptions of typical hypertext systems do not apply. The novelty of the Web leads to new problems in information access, and it is necessary to make use of the new kinds of information available, such as multiple independent categorization, naming, and indexing of pages. This paper discusses the varieties of link information (not just hyperlinks) on the Web, how the Web differs from conventional hypertext, and how the links can be exploited to build useful applications. Specific applications presented as part of the ParaSite system find individuals' homepages, new locations of moved pages, and unindexed information.

Introduction

The World-Wide Web contains millions of pages of data. Practical access to this information requires applying and expanding hypertext research to build powerful search tools. Most Web search tools only make use of the text on a page, ignoring another rich source of information, the links among pages. Much human thought has gone into creating each hyperlink and labeling it with anchor text. Other valuable relational information can be gleaned from the structure, hierarchy, and similarity of pieces of text. This information is already used by individuals when they browse the Web. It should be harnessed to build powerful automatic search tools.

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

Wayback Replay

F11 2019



fatcat! About Guide Changelog Search Papers... bnewbold-archive

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) Metadata Edit History

Abstract

<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-to-read 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>

inapplication/xml+jats format

► Published in [PeerJ](#) by PeerJ

Known Files and URLs

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

Read Full Text

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

DOI [10.7717/peerj.4375](#)
PubMed [29456894](#)
PMC [PMC5815332](#)
Wikidata [Q49873702](#)

Container Metadata
Open Access Publication
In DOAJ
In ISSN ROAD
ISSN-L: 2167-8359
Fatcat Entry

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

Edit Metadata View History

Fulltext
Mirror
Locations

Publication Status

PIDs Galore

Journal Metadata

Alt. Versions

API Helpers

Wiki Sauce

F11 2019



Editing Metadata

Edit Container Entity

See the [catalog style guide](#) for schema notes, and the [editing tutorial](#) if this is your first time making an edit.

Editgroup Metadata

Select an in-progress editgroup for this change to be part of (or start a new one):

editgroup_yf2ydu5fbnv1tqbg123b5gsm 2019-10-10 09:04:11.363759+00:00
small random edits

If starting a new editgroup, you can add a description for the whole group:

Editgroup Description

The Basics

Container Type: Scholarly Journal

Name/Title: Journal of Indian Society of Pedodontics and Preventive Dentistry

Original Name (native language)

Publisher: Medknow Publications

Country of Publication (ISO code): in

ISSN (linking): 0970-4388

ISSN (electronic): 1998-3905

ISSN (print): 0970-4388

Wikidata QID: Q6295327

Homepage URLs

Landing page or mirror locations of container as a whole:

1 http://www.jsppd.com/

Submit

Description of Changes

This description will be attached to the individual edit, and to the editgroup as a whole.

Update Container!

Edit will be part of the current editgroup, which needs to be submitted and approved before the change is included in the catalog.

Editgroup

editgroup_yf2ydu5fbnv1tqbg123b5gsm

What is an editgroup? An editgroup is a set of entity edits, bundled together into a coherent, reviewable bundle.

Status: Not Submitted
Editor: bnewbold-archive
Description: small random edits

Accept Edits
Submit

Release Edits (0)

Work Edits (0)

Container Edits (1)

container_zrxnp7o3mzho5o2pqborropua updated [view edit] [re-edit] [delete-edit]
Revision: 7db46f25-c4c9-4a2d-855b-5b0dc88fe1a0

description url; title formatting

Creator Edits (0)

File Edits (0)

File Set Edits (0)

Web Capture Edits (0)

Comments and Annotations

bnewbold-archive Admin at 2019-10-11 01:34:12

This is an example comment on an editgroup. Hooray review!

Add Comment

Markdown is allowed

Submit

Recent Changes

change log

Limited to the most recent entries.

Changelog Index	Editgroup	Description
#2976548 2019-10-11 01:31:57	crossref-bot 75a64ub3h5cy7oq2jeshafe1r4	Automated import of Crossref DOI metadata, harvested from REST API
#2976547 2019-10-11 01:31:56	crossref-bot gfr1q5lwufettkpa1edn12c4u	Automated import of Crossref DOI metadata, harvested from REST API
#2976546 2019-10-11 01:31:55	crossref-bot 1o3mhmwx4jczu01uy6f4172spq	Automated import of Crossref DOI metadata, harvested from REST API
#2976545 2019-10-11 01:31:54	crossref-bot y1xqn1zhyfhs3rjb0be5vhtqbaa	Automated import of Crossref DOI metadata, harvested from REST API
#2976544 2019-10-11 01:31:52	crossref-bot 14femp2ab7boxfr547x12ts24i	Automated import of Crossref DOI metadata, harvested from REST API
#2976543 2019-10-11 01:31:51	crossref-bot 31vvugcdqnb3arcgcf6umv1uy	Automated import of Crossref DOI metadata, harvested from REST API
#2976542 2019-10-11 01:31:50	crossref-bot qe2vb7qevng1fcsr2q8q5e273a	Automated import of Crossref DOI metadata, harvested from REST API
#2976541 2019-10-11 01:31:49	crossref-bot ap7amyk3cncu3fep41y1u5xgrm	Automated import of Crossref DOI metadata, harvested from REST API
#2976540 2019-10-11 01:31:48	crossref-bot xuw24q4p7fedfmbh1gy2pnrty	Automated import of Crossref DOI metadata, harvested from REST API
#2976539 2019-10-11	crossref-bot unpnc3v0v1kx31dnt131om	Automated import of Crossref DOI metadata, harvested from REST API

Edit

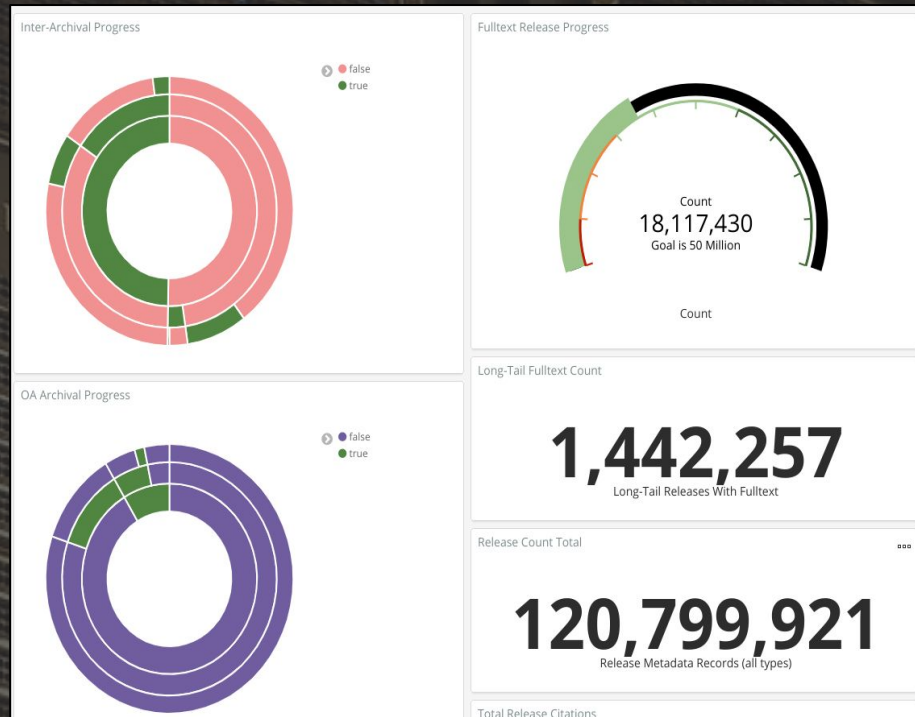
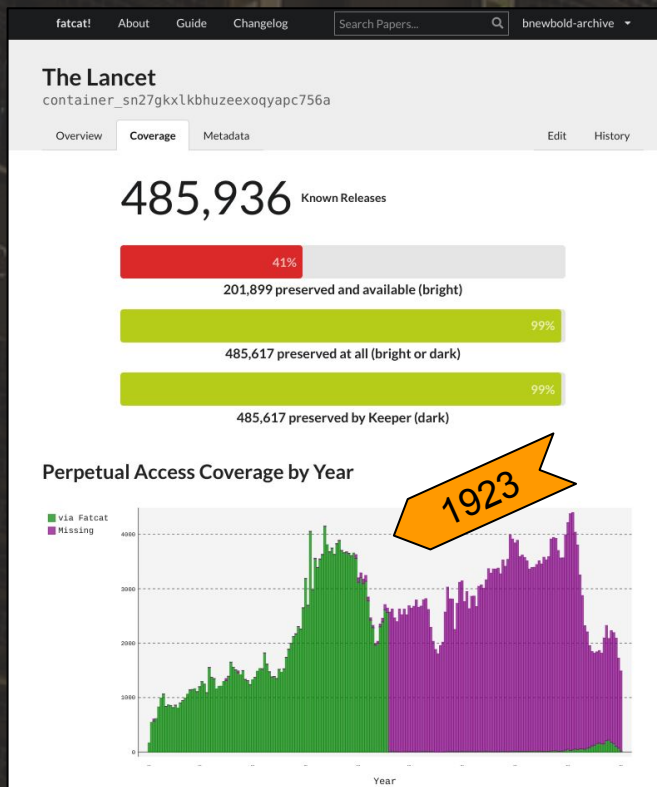
Collaborative Review

Changelog

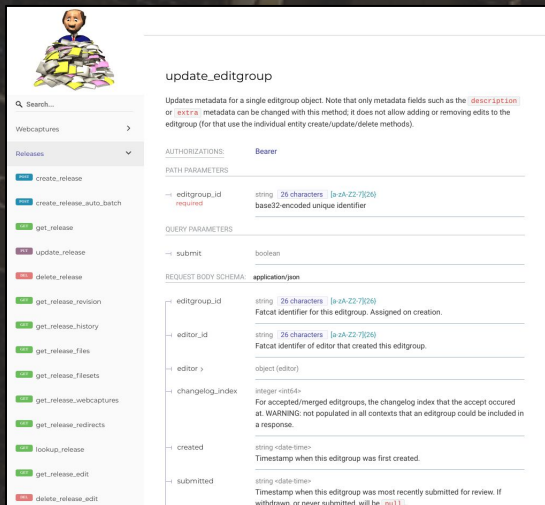
F11 2019



Coverage Dashboards



#DeveloperThings



update_editgroup

Updates metadata for a single editgroup object. Note that only metadata fields such as the `description` or `notes` metadata can be changed with this method; it does not allow adding or removing edits to the editgroup (for that use the individual entity create/update/delete methods).

Webcaptures

- create_release
- create_release_auto_batch
- get_release
- update_release
- delete_release
- get_release_revision
- get_release_history
- get_release_files
- get_release_files
- get_release_webcaptures
- get_release_redirects
- lookup_release
- get_release_edit
- delete_release_edit

AUTHORIZATIONS

Bearer

PATH PARAMETERS

Parameter	Type	Description
editgroup_id	string (26 characters) [a-zA-Z0-9]	base32-encoded unique identifier

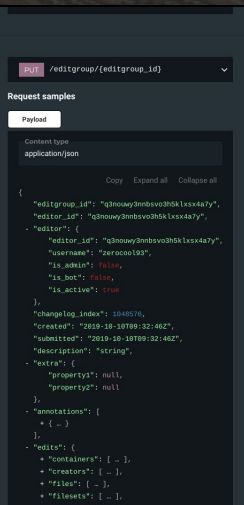
QUERY PARAMETERS

Parameter	Type	Description
submit	boolean	

REQUEST BODY SCHEMA

application/json

Field	Type	Description
editgroup_id	string (26 characters) [a-zA-Z0-9]	Fatcat identifier for this editgroup. Assigned on creation.
editor_id	string (26 characters) [a-zA-Z0-9]	Fatcat identifier of editor that created this editgroup.
editor	object (editor)	
changelog_index	integer (uint8)	For accepted/merged editgroups, the changelog index that the accept occurred at. WARNING: not populated in all contexts that an editgroup could be included in a response.
created	string (date-time)	Timestamp when this editgroup was first created.
submitted	string (date-time)	Timestamp when this editgroup was most recently submitted for review. If withdrawn, or never submitted will be <code>null</code> .

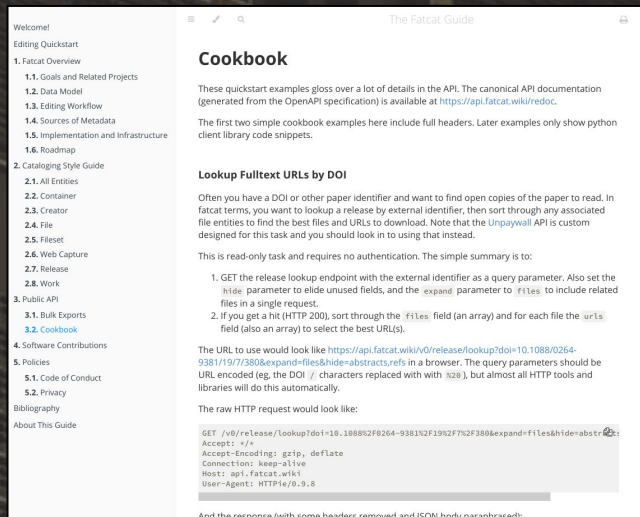


Request samples

Payload

Content type: application/json

```
{
  "editgroup_id": "3h0uwy3mbvsv30k1ss447y",
  "editor_id": "qhm0uwy3mbvsv30k1ss447y",
  "editor": {
    "editor_id": "qhm0uwy3mbvsv30k1ss447y",
    "username": "zerocoin123",
    "is_admin": false,
    "is_bot": false,
    "is_active": true
  },
  "changelog_index": 1048070,
  "created": "2019-10-10T09:32:46Z",
  "submitted": "2019-10-10T09:32:46Z",
  "description": "string",
  "extra": {
    "properties": null,
    "property": null
  },
  "annotations": [
    { }
  ],
  "edits": [
    {
      "containers": [ ],
      "creators": [ ],
      "files": [ ],
      "filesets": [ ]
    }
  ]
}
```



Welcomel

Editing Quickstart

1. Fatcat Overview

1.1. Goals and Related Projects

1.2. Data Model

1.3. Editing Workflow

1.4. Sources of Metadata

1.5. Implementation and Infrastructure

1.6. Roadmap

2. Cataloging Style Guide

2.1. All Entities

2.2. Container

2.3. Creator

2.4. File

2.5. Fileset

2.6. Web Capture

2.7. Release

2.8. Work

3. Public API

3.1. Bulk Exports

3.2. Cookbook

4. Software Contributions

5. Policies

5.1. Code of Conduct

5.2. Privacy

Bibliography

About This Guide

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 entries to find the best files and URLs to download. Note that the `Unpaywall` 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 best URL(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>

The query parameters should be URL encoded (eg. the DOI `//` characters replaced with `%2F`), but almost all HTTP tools and libraries will do this automatically.

The raw HTTP request would look like:

```
GET /v0/release/lookup?doi=10.1088/0264-9381/19/7/380&expand=files&hide=abstracts
Accept: */*
Accept-Encoding: gzip, deflate
Connection: keep-alive
Host: api.fatcat.wiki
User-Agent: HTTPie/0.9.8
```

And the response (with some headers removed and JSON body unescaped):

API Specification and Docs

<https://api.fatcat.wiki>

“The Guide”

<https://guide.fatcat.wiki>

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

F11 2019



Outline

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

PARTNER

- 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



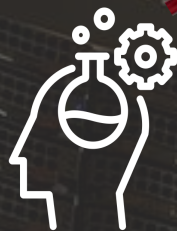
F11 2019

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**

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

ARCHIVE



Created by Chameleon Design
from Noun Project



Created by ProSymbols
from Noun Project



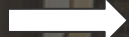
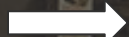
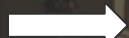
Created by ProSymbols
from Noun Project



Created by ProSymbols
from Noun Project



Created by ProSymbols
from Noun Project



Created by Pham Duy Phuong Hung
from Noun Project

PARTNER



F11 2019

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

