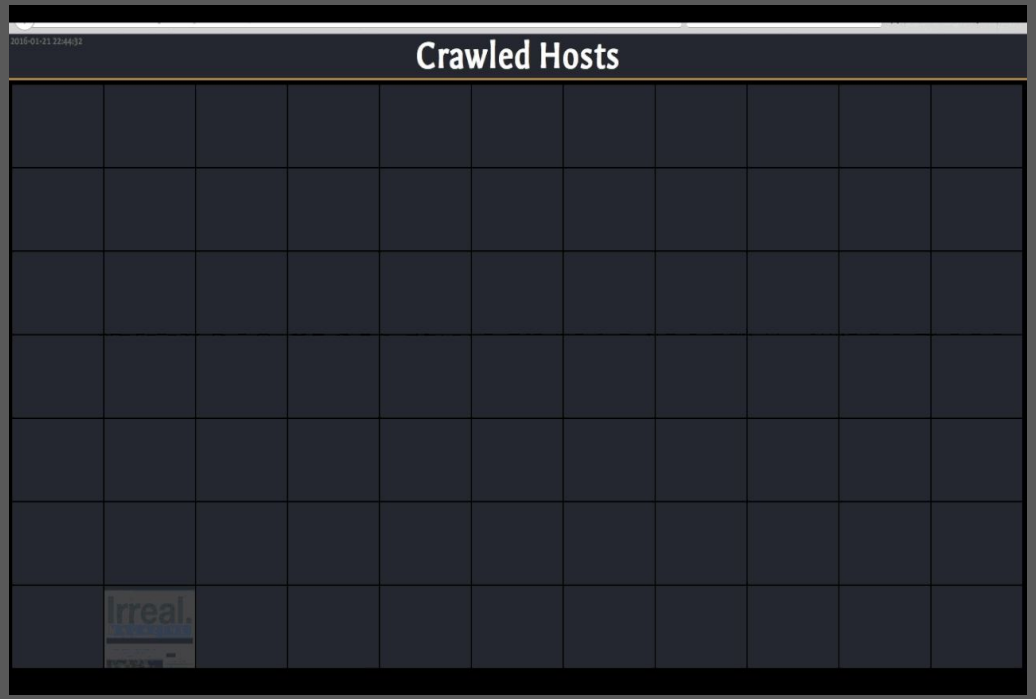
Perpetual Access Machines:
Archiving Web-published Scholarship at Scale



≠

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

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)

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

F11 2019

Outline

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

Outline

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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**

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

Methods / Approaches

Forever Approach

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

Outline

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

INTERNET ARCHIVE

NEW!

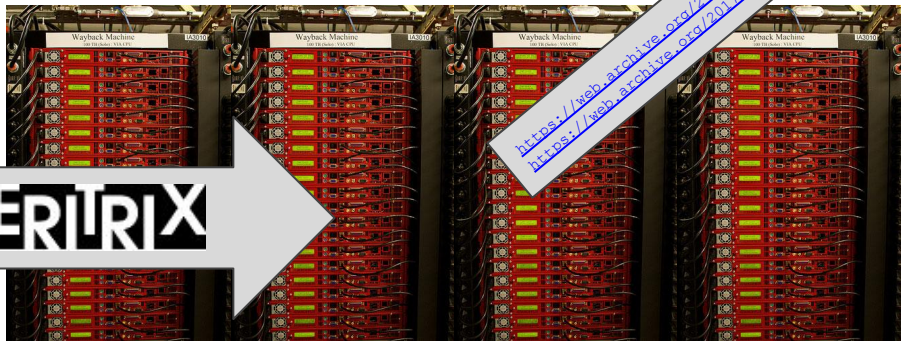
"fatcat"

NEW!

"sandcrawler"

WaybackMachine

https://web.archive.org/20190102O3/http://plos.org/...
https://web.archive.org/20170102O3/http://elife.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

CiteSeerX

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

https://archive.org/details/ia_biblio_metadata

82 RESULTS

SORT BY VIEWS · TITLE · DATE

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

- 2
- 1
- 1
- 1
- 1
- 1

Crossref DOI Dump (2018-01)
459 1 0

Crossref DOI Dump (2018-09)
259 0 0

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

oaDOI DOI/URL Dataset
226 0 0

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

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

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

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

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

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

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

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

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

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

Journal ISSN Metadata Exploration (2018-04-05)
by Internet Archive Web Group

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**

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

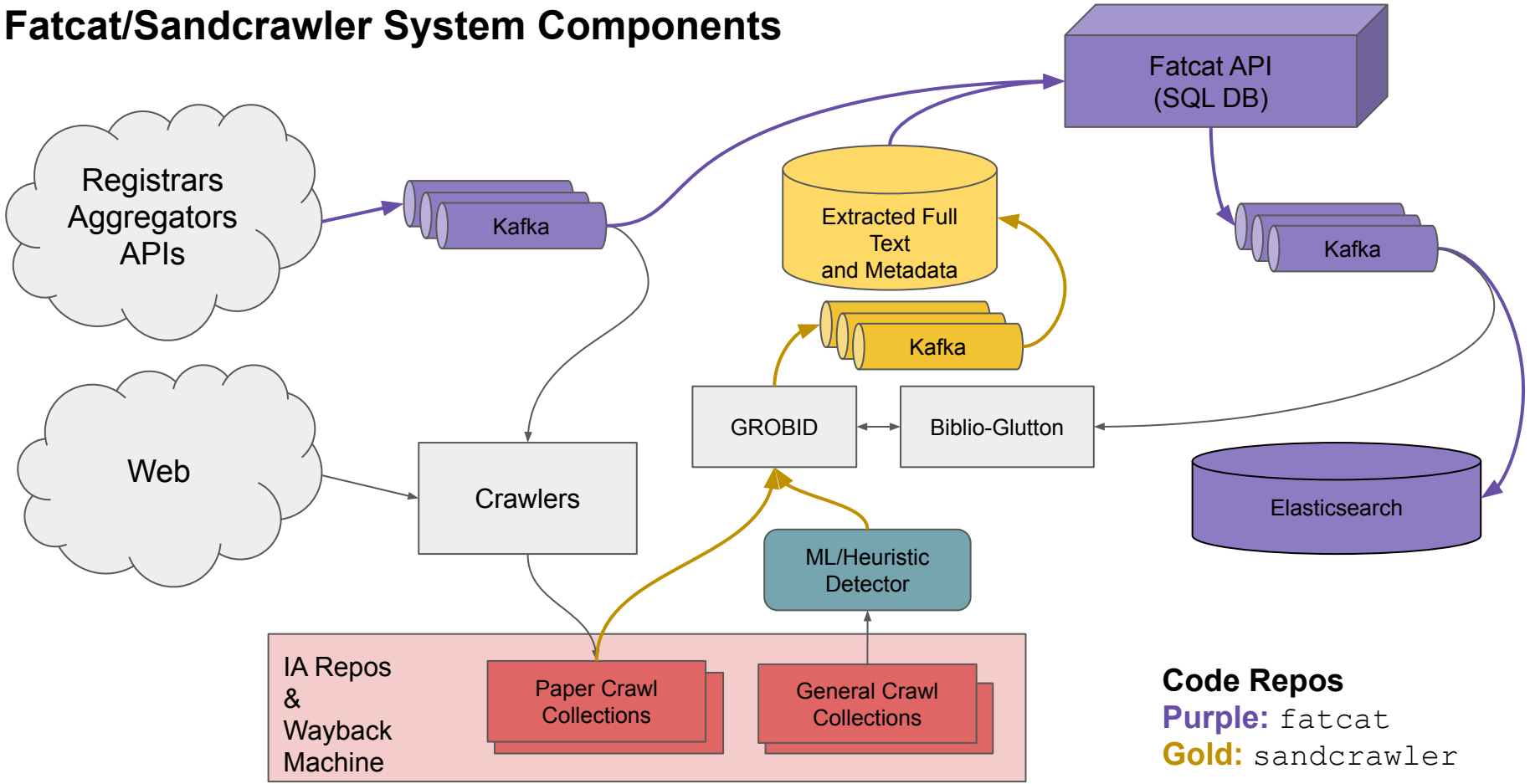
```
PS3557
.R5355          Grisham, John
F57 1991

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

          1. Government investigators--Fiction.
          2. Organized crime--Fiction.
```

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

Fatcat/Sandcrawler System Components



Outline

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

https://fatcat.wiki/

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

by title, authors, identifiers... Search

96,947,165	18,117,429	140,085
Papers	Fulltext	Journals

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

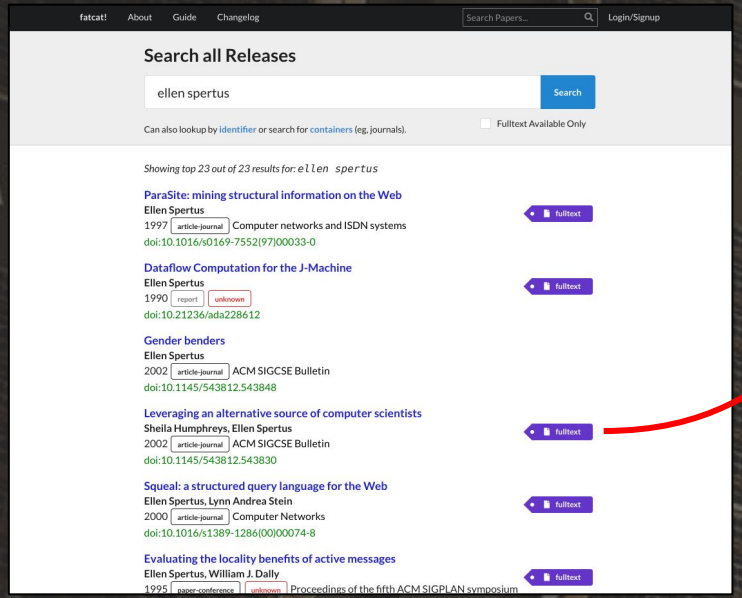
Features include archival file-level metadata (verified digests and long-term copies), an open, documented API, and work/release indexing (eg, distinguishing between and linking pre-prints, manuscripts, and version-of-record). Read more...

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

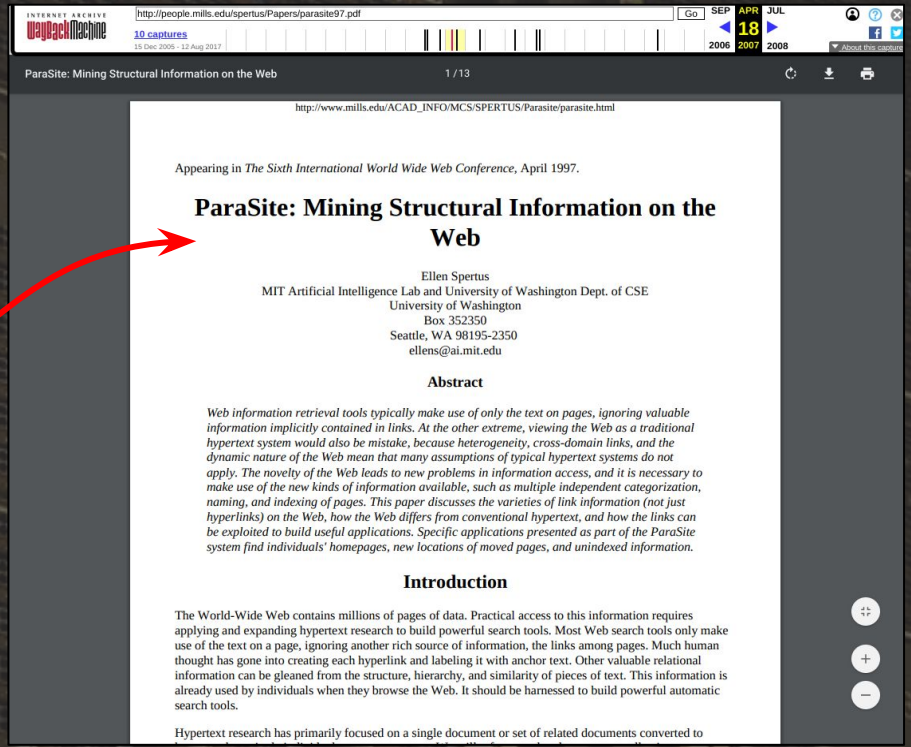
Development funding comes from The Andrew Mellon Foundation to improve preservation and access to "long-tail" open access works on the public web which might otherwise be lost.

INTERNET ARCHIVE

Basic Reader Access



Search



Wayback Replay

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

Publication Status

PIDs Galore

Journal Metadata

Alt. Versions

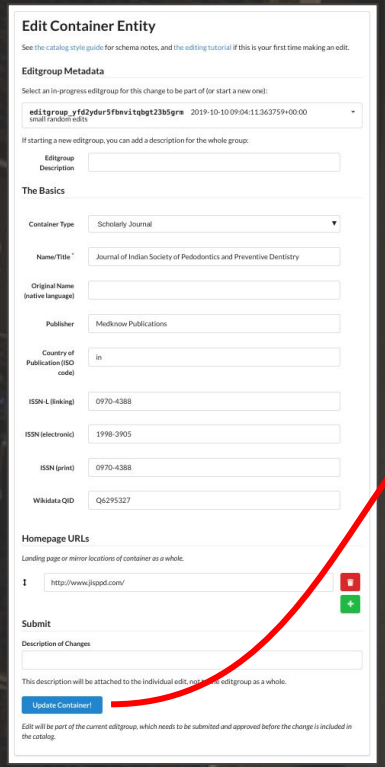
API Helpers

Wiki Sauce

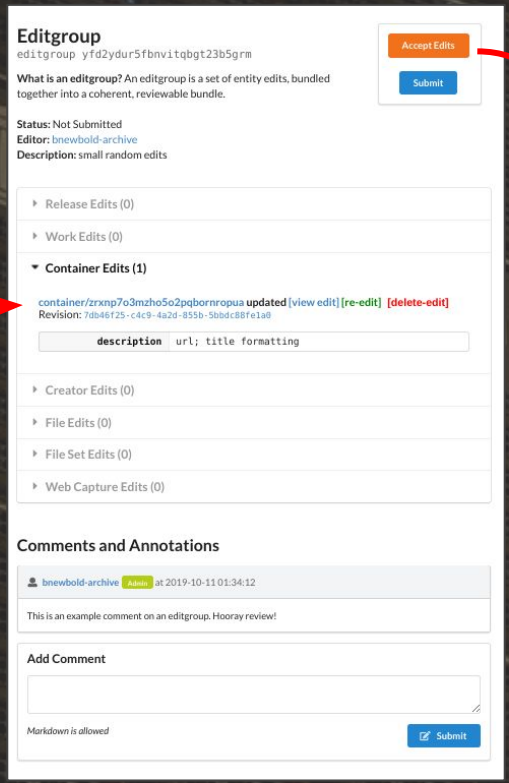
Fulltext Mirror Locations

F11 2019

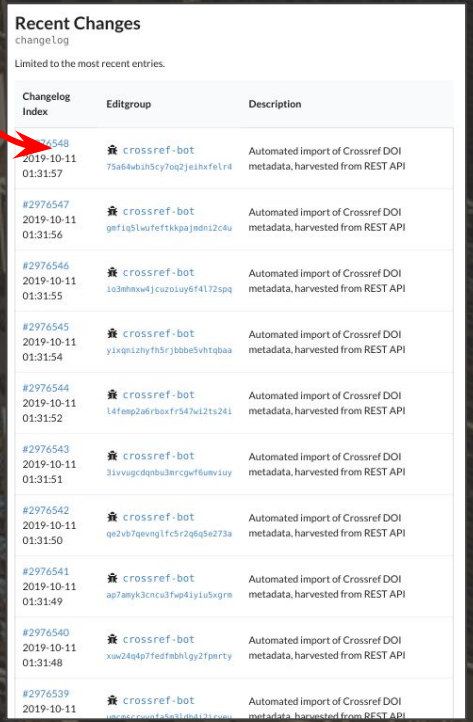
Editing Metadata



Edit



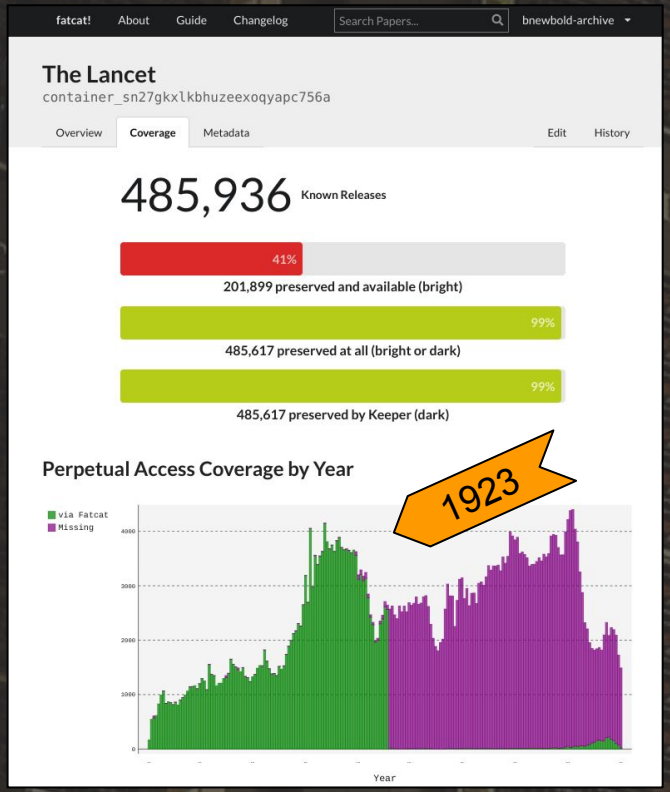
Collaborative Review

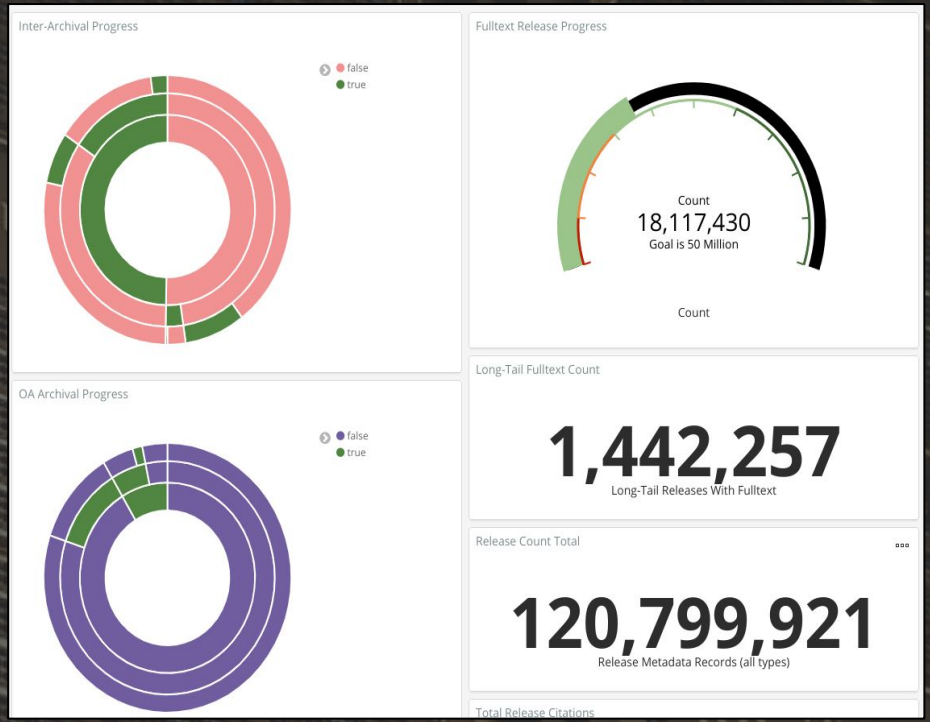


Changelog

F11 2019

Coverage Dashboards





Journal Level

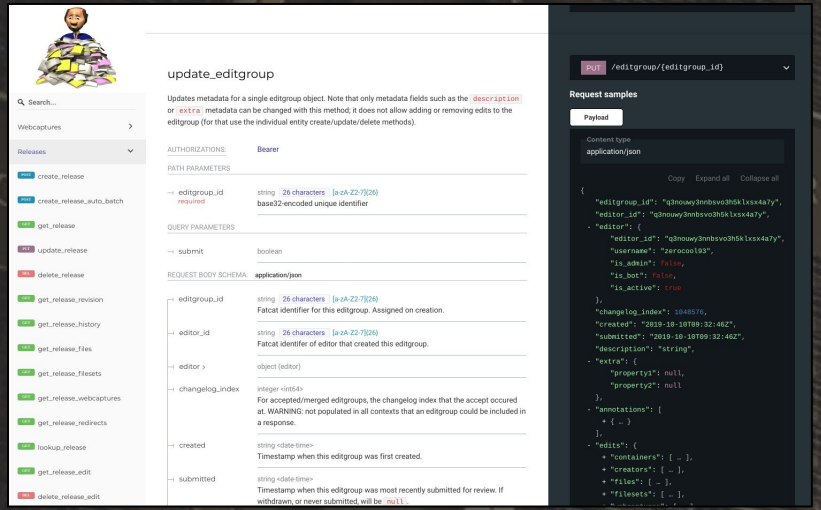
Collection Level

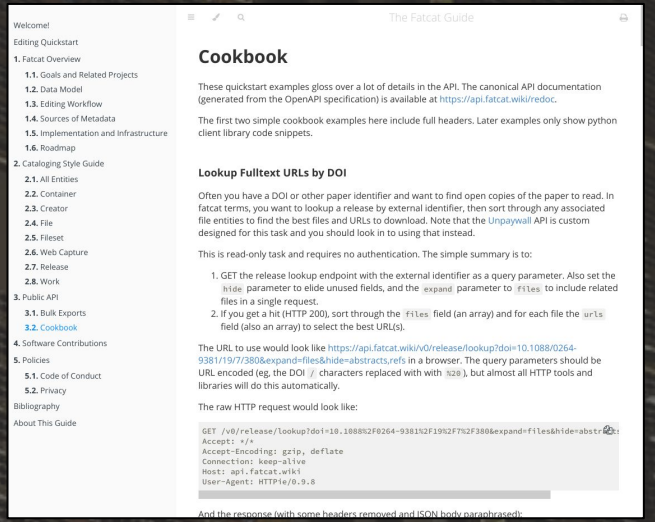
INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

#DeveloperThings





API Specification and Docs

https://api.fatcat.wiki

“The Guide”

https://guide.fatcat.wiki

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

Outline

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

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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

THE ANDREW W. MELLON FOUNDATION

Semantic Scholar

COS CENTER FOR OPEN SCIENCE

our research

University of California CDL California Digital Library

LOCKSS

INSTITUTE of Museum and Library SERVICES

INTERNET ARCHIVE

ARCHIVE-IT

dat://

ISSN

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

INTERNET ARCHIVE

F11 2019

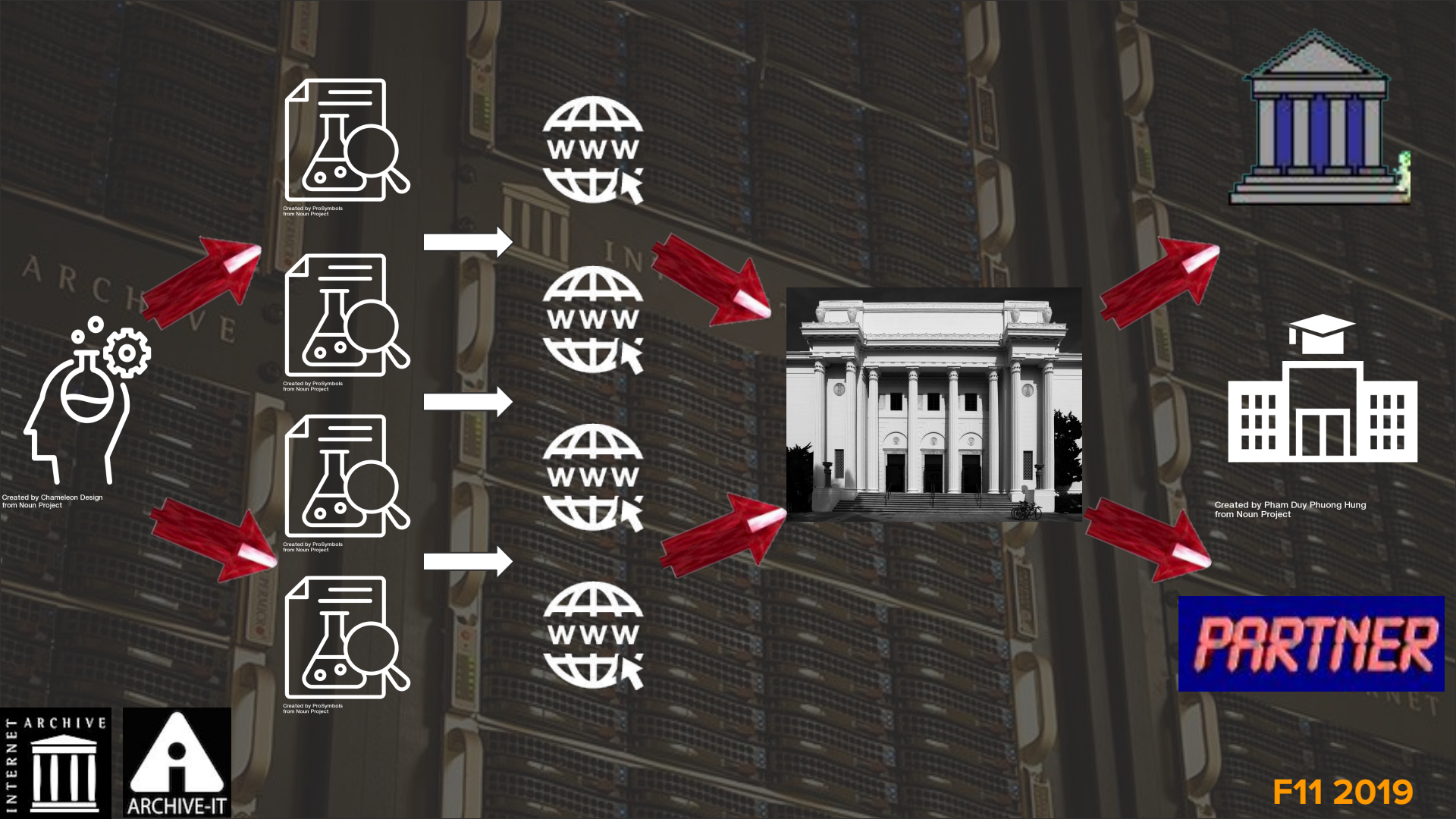
Current Work & Roadmap

- 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

https://fatcat.wiki/stats

F11 2019


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 Chameleon Design from Noun Project
Created by Pham Duy Phuong Hung from Noun Project
PARTNER
INTERNET ARCHIVE
ARCHIVE-IT
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!

INTERNET ARCHIVE

ARCHIVE-IT

F11 2019

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, https://archive.org
Fatcat beta, https://fatcat.wiki/

F11 2019