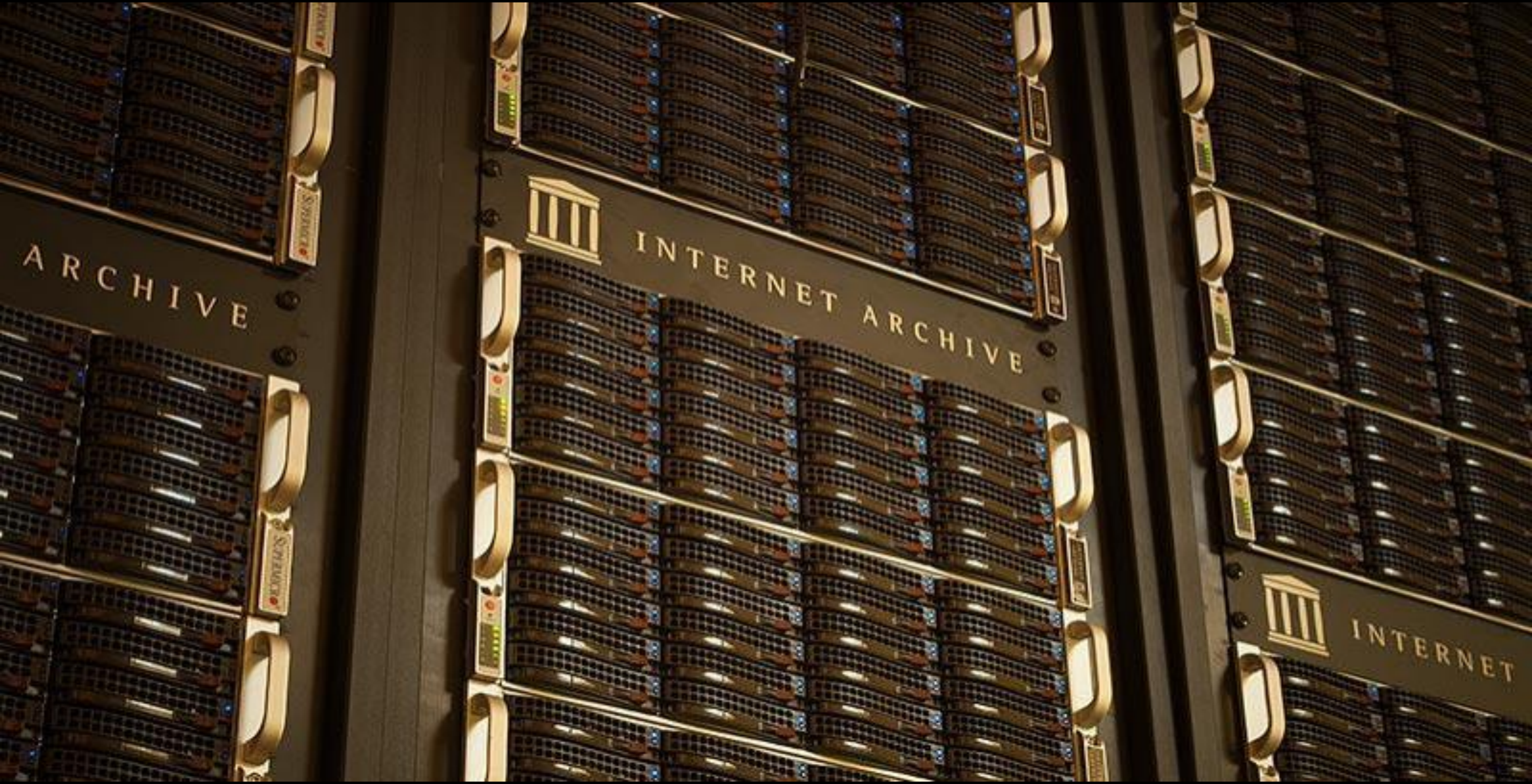
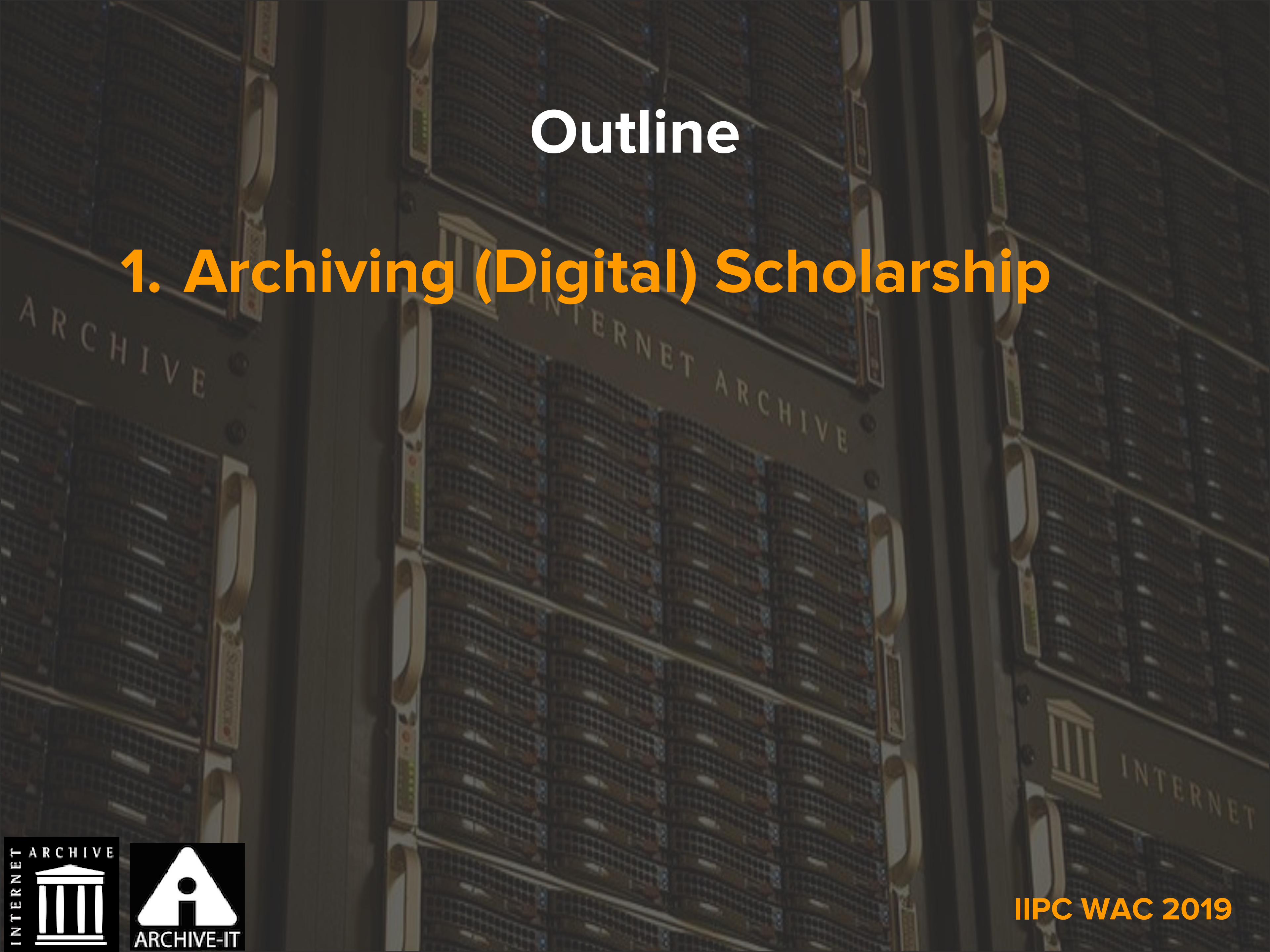
From Open Access to Perpetual Access: Archiving Web-Published Scholarship



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

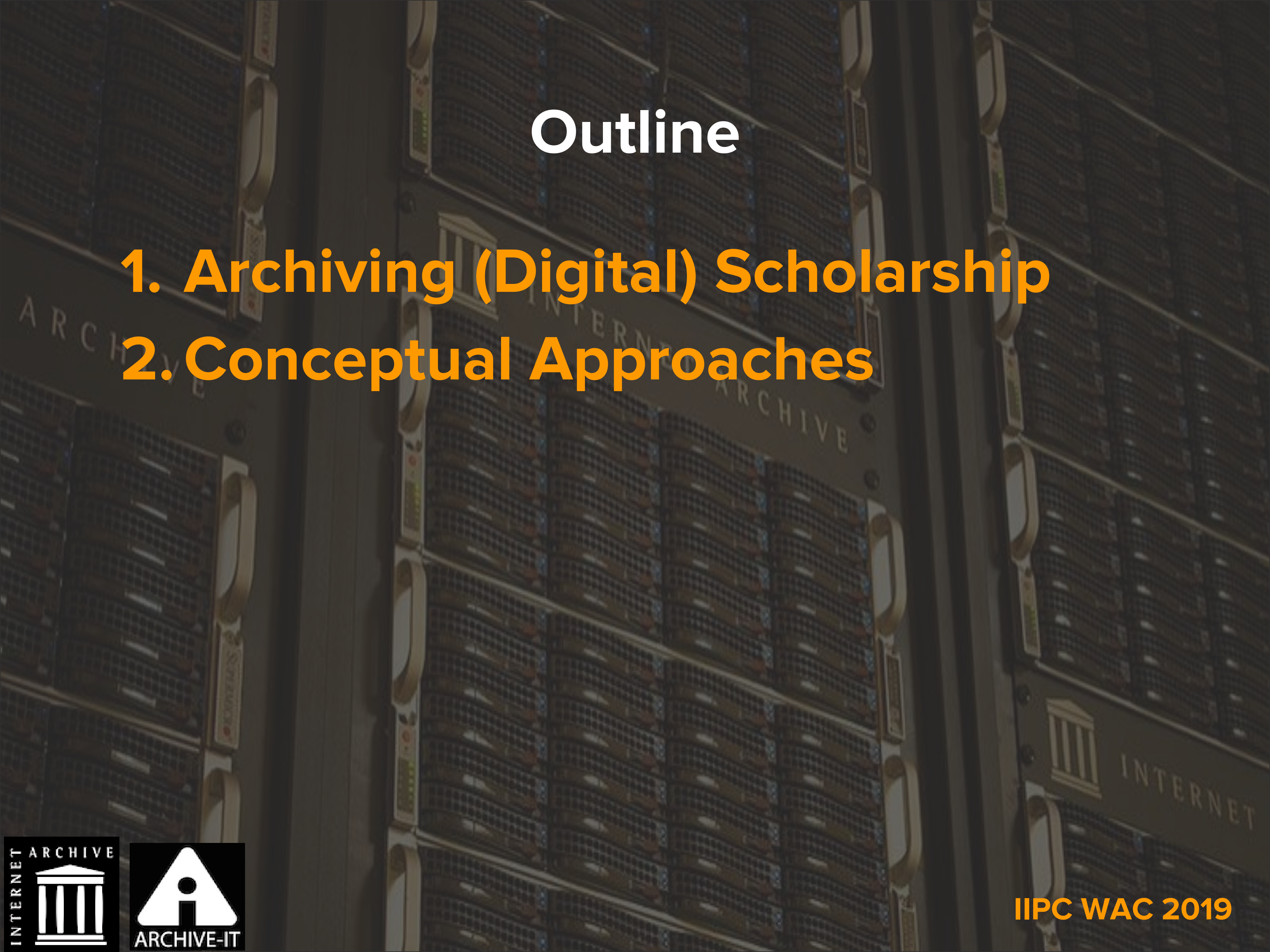
Build a complete, use-oriented, highly-available archive and knowledge graph of every publicly-accessible scholarly output + descriptive metadata and full-text, linked with versions and secondary outputs (data/blogs/etc) with a priority on long-tail, at-risk publications -- all accessible via API-first 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

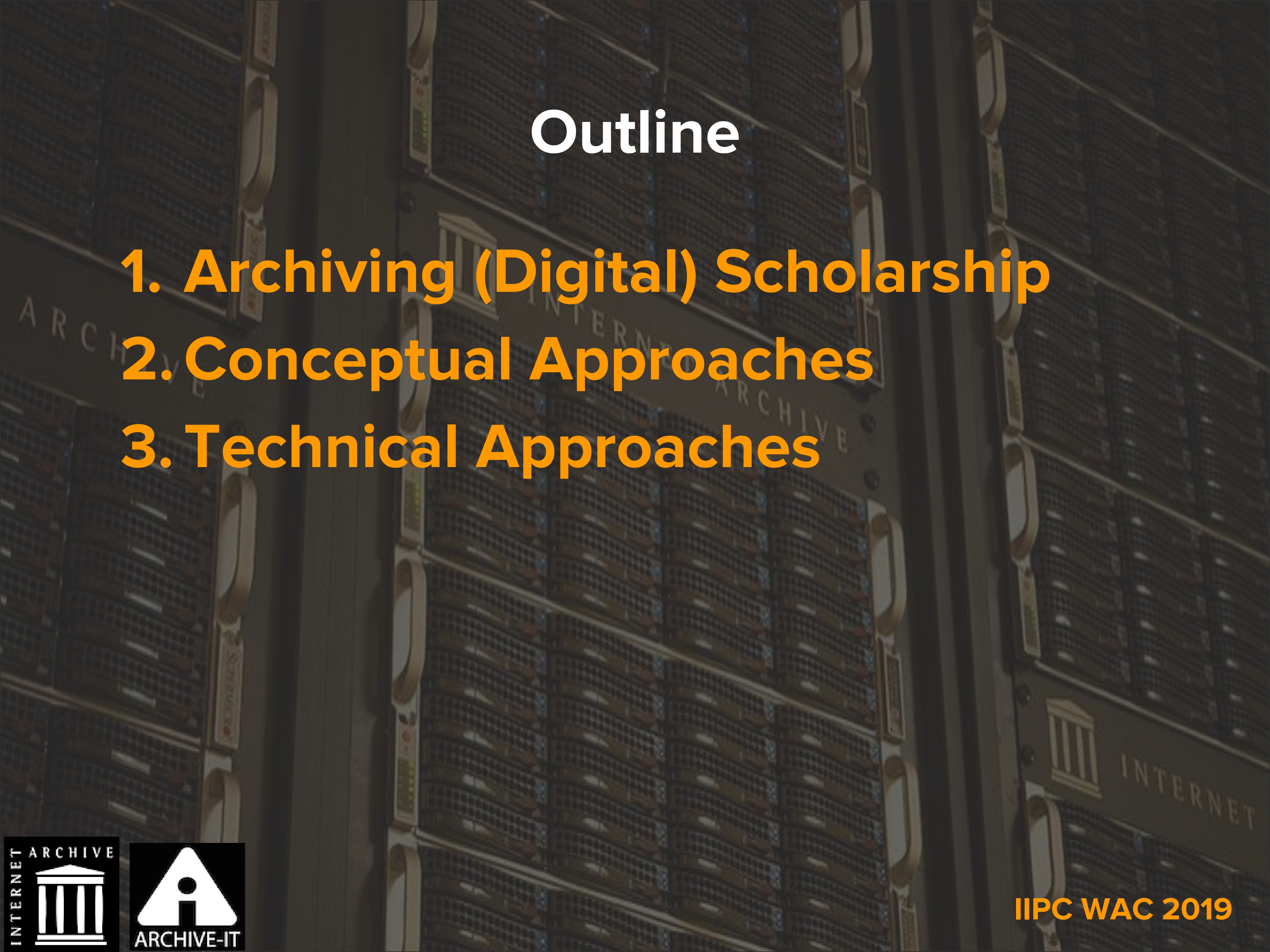
1. Archiving (Digital) Scholarship
2. Conceptual Approaches

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

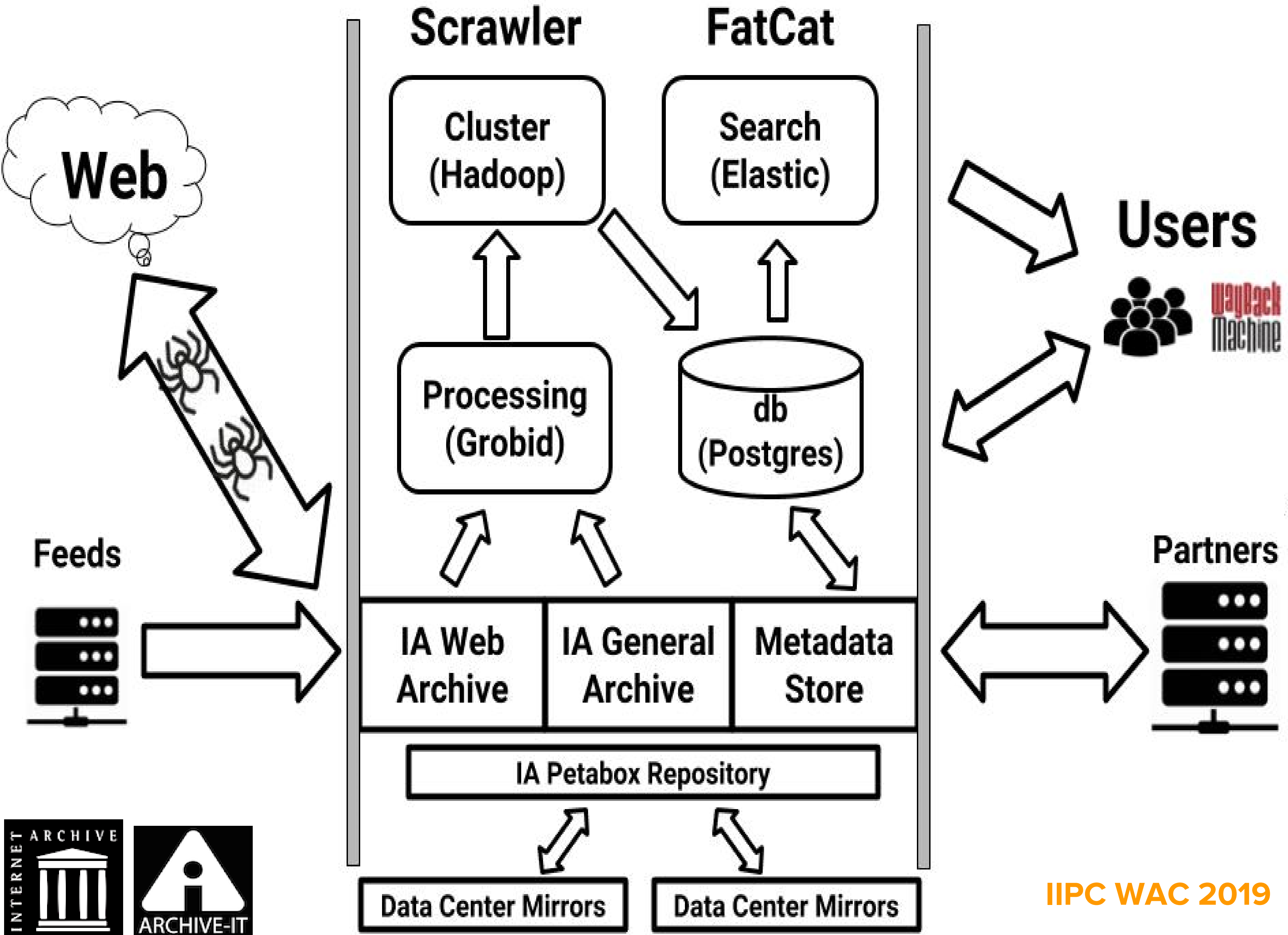
Conceptual Approaches 2

- 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



Web

Scrawler

FatCat

Cluster (Hadoop)

Search (Elastic)

Users

WayBack Machine

Processing (Grobid)

db (Postgres)

Feeds

Partners

IA Web Archive

IA General Archive

Metadata Store

IA Petabox Repository

Data Center Mirrors

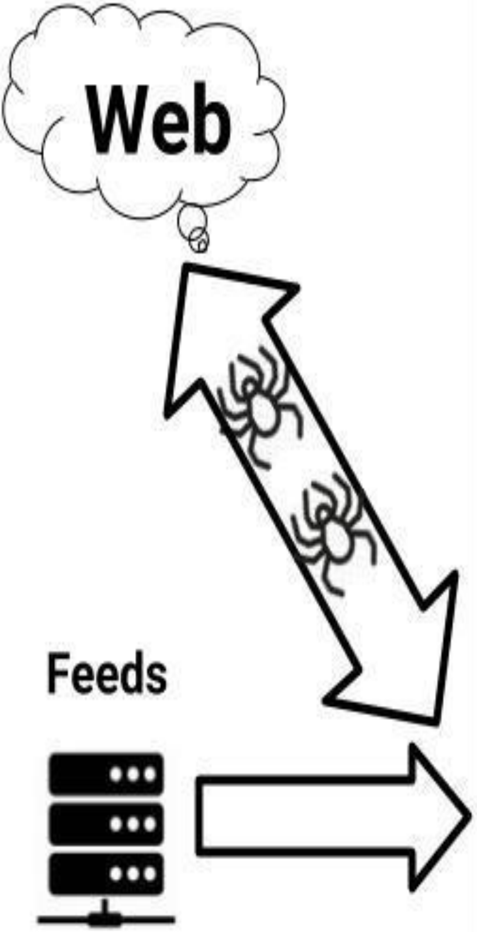
Data Center Mirrors

INTERNET ARCHIVE

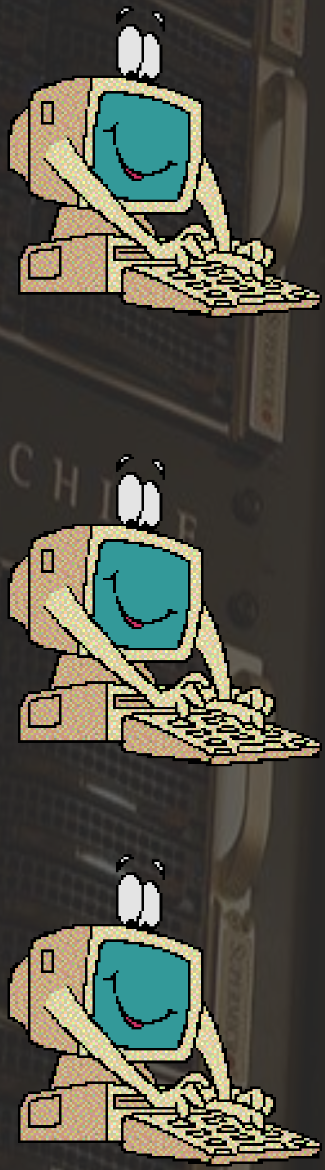
ARCHIVE-IT

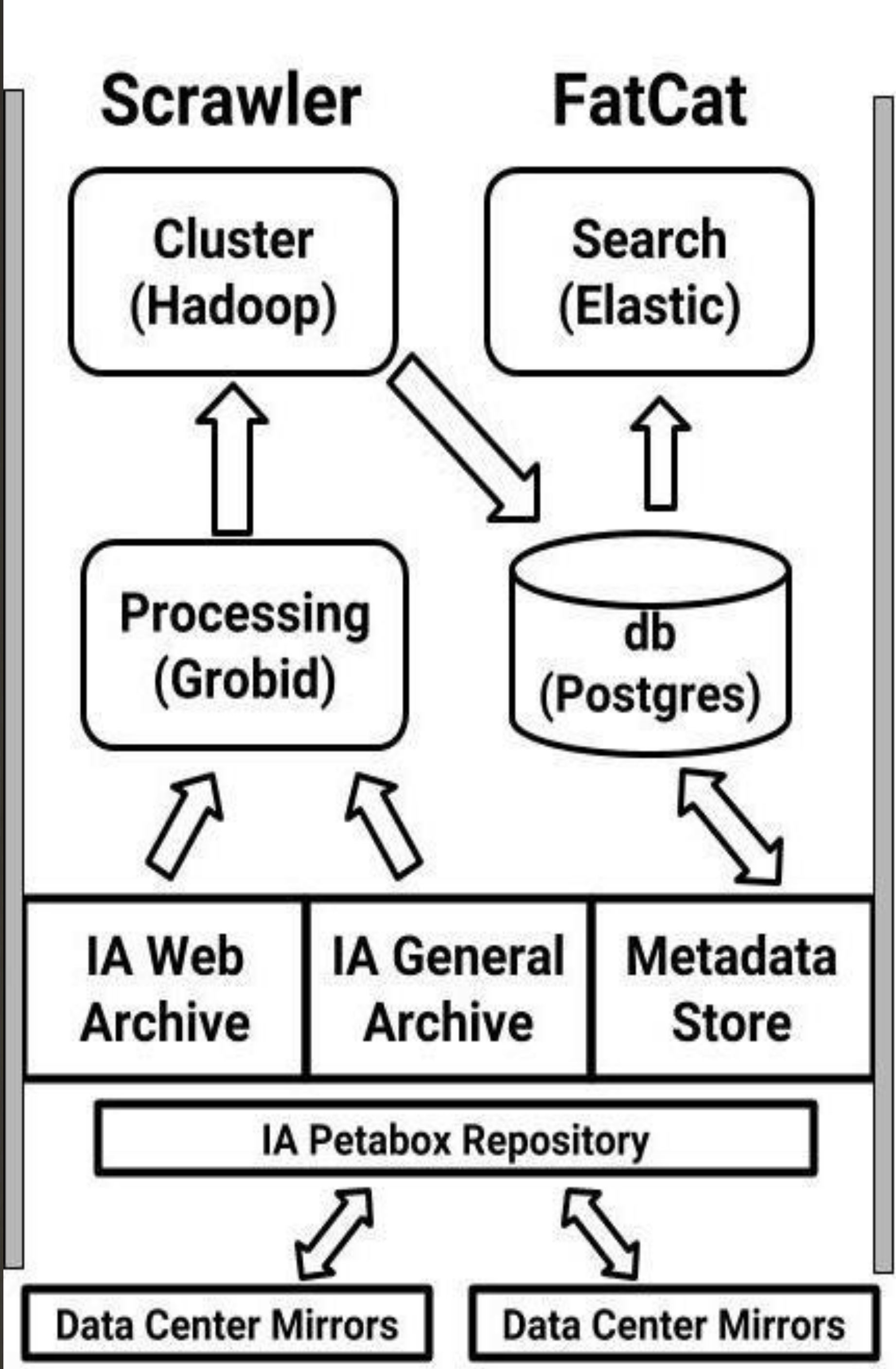
IIPC WAC 2019

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)





Scrawler

FatCat

Cluster (Hadoop)

Search (Elastic)

Processing (Grobid)

db (Postgres)

IA Web Archive

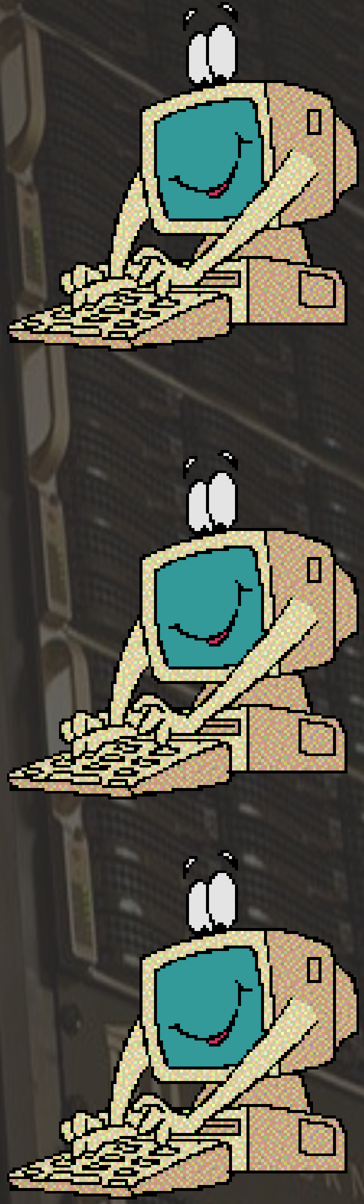
IA General Archive

Metadata Store

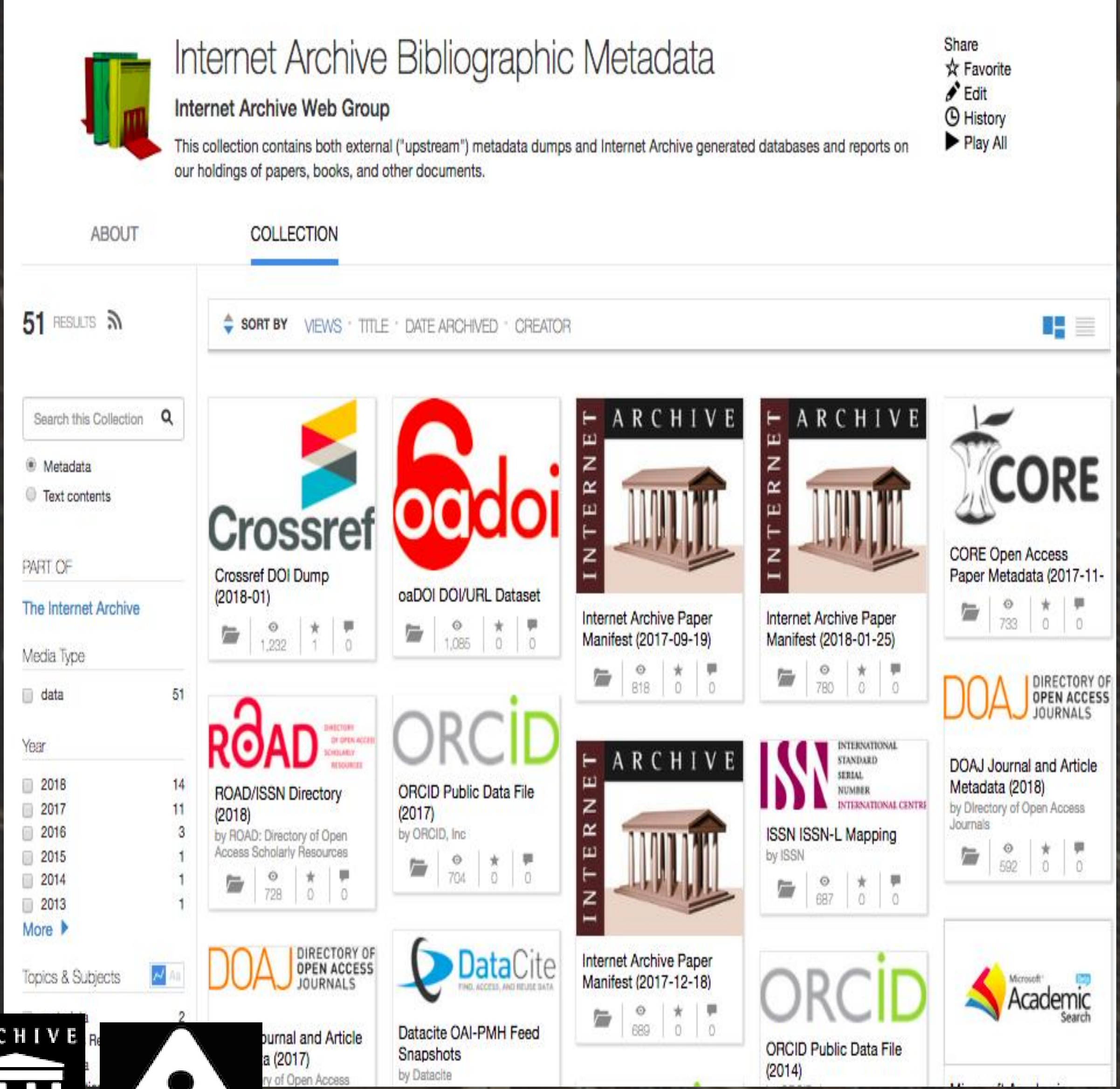
IA Petabox Repository

Data Center Mirrors

Data Center Mirrors



Sources Archived



Internet Archive 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
Edit
History
▶ Play All

ABOUT COLLECTION

51 RESULTS

SORT BY VIEWS · TITLE · DATE ARCHIVED · CREATOR

Search this Collection

● Metadata
○ Text contents

PART OF

The Internet Archive

Media Type

□ data	51

Year

□ 2018	14
□ 2017	11
□ 2016	3
□ 2015	1
□ 2014	1
□ 2013	1

More ▸

Topics & Subjects

Crossref DOI Dump (2018-01)
1,232 1 0

oaDOI DOI/URL Dataset
1,085 0 0

Internet Archive Paper Manifest (2017-09-19)
818 0 0

Internet Archive Paper Manifest (2018-01-25)
780 0 0

CORE Open Access Paper Metadata (2017-11-
733 0 0

ROAD/ISSN Directory (2018)
by ROAD: Directory of Open Access Scholarly Resources
728 0 0

ORCID Public Data File (2017)
by ORCID, Inc
704 0 0

Internet Archive Paper Manifest (2017-12-18)
689 0 0

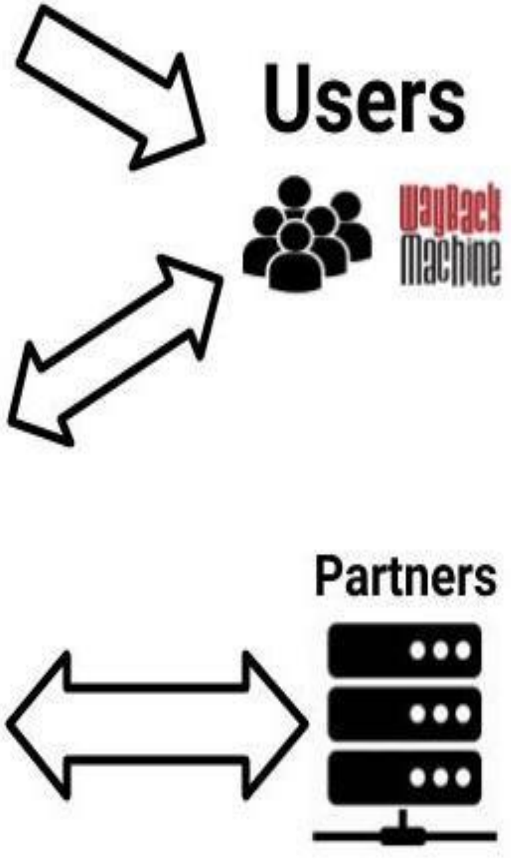
ISSN ISSN-L Mapping
by ISSN
687 0 0

DOAJ Journal and Article Metadata (2018)
by Directory of Open Access Journals
592 0 0

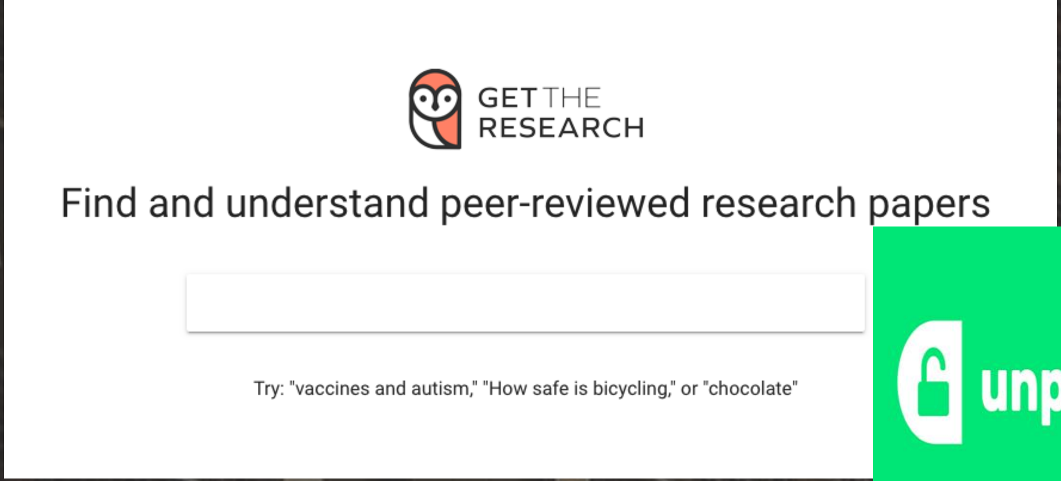
...urnal and Article ...a (2017)
...y of Open Access

Datacite OAI-PHM Feed Snapshots
by Datacite

ORCID Public Data File (2014)



Partnerships



GET THE RESEARCH

Find and understand peer-reviewed research papers

Try: "vaccines and autism," "How safe is bicycling," or "chocolate"

unpaywall

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INTERNATIONAL CENTRE

WIKIDATA

Semantic Scholar

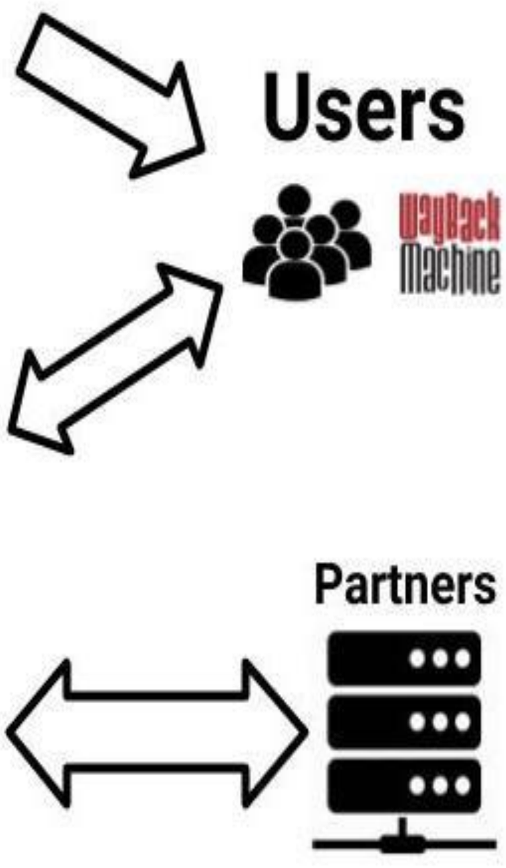
OpenCitations

DOAJ
DIRECTORY OF
OPEN ACCESS
JOURNALS

Crossref

INTERNET ARCHIVE

ARCHIVE-IT

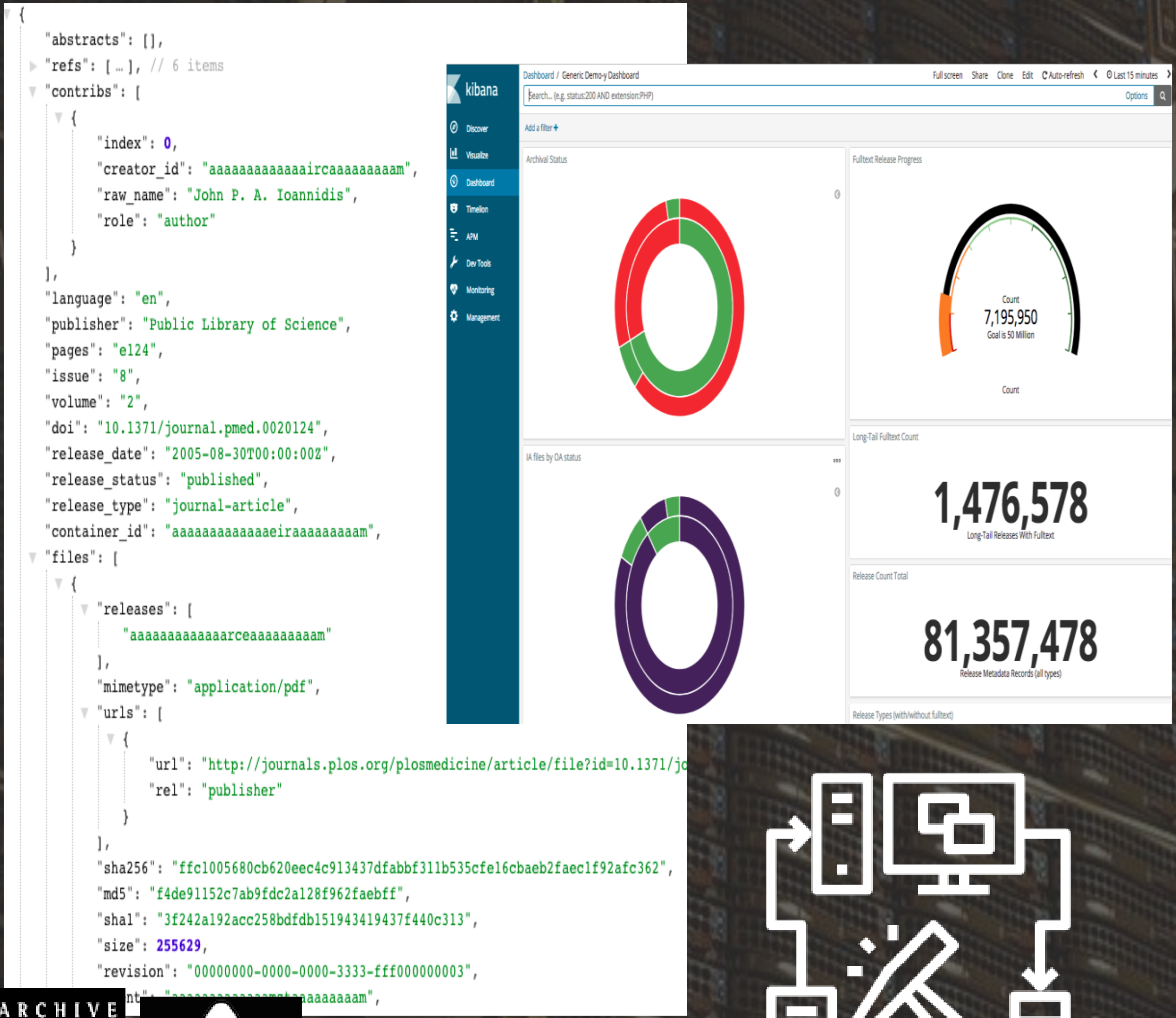


Users

Wayback Machine

Partners

APIs, Reporting, Bulk Access



```
▾ {
    "abstracts": [],
  ▸ "refs": [ … ], // 6 items
  ▾ "contribs": [
      ▾ {
            "index": 0,
            "creator_id": "aaaaaaaaaaaaaircaaaaaaaaaam",
            "raw_name": "John P. A. Ioannidis",
            "role": "author"
        }
    ],
    "language": "en",
    "publisher": "Public Library of Science",
    "pages": "e124",
    "issue": "8",
    "volume": "2",
    "doi": "10.1371/journal.pmed.0020124",
    "release_date": "2005-08-30T00:00:00Z",
    "release_status": "published",
    "release_type": "journal-article",
    "container_id": "aaaaaaaaaaaaaeiraaaaaaaaam",
  ▾ "files": [
      ▾ {
          ▾ "releases": [
                "aaaaaaaaaaaaarceaaaaaaaaaam"
            ],
            "mimetype": "application/pdf",
          ▾ "urls": [
              ▾ {
                    "url": "http://journals.plos.org/plosmedicine/article/file?id=10.1371/jo
                    "rel": "publisher"
                }
            ],
            "sha256": "ffc1005680cb620eec4c913437dfabbf311b535cfe16cbaeb2faec1f92afc362",
            "md5": "f4de91152c7ab9fdc2a128f962faebff",
            "sha1": "3f242a192acc258bdfdb151943419437f440c313",
            "size": 255629,
            "revision": "00000000-0000-0000-3333-fff000000003",
           nt": "aaaaaaaaaaaaamztaaaaaaaaam",
```

Dashboard / Generic Demo-y Dashboard

Search... (e.g. status:200 AND extension:PHP)

Archival Status

Fulltext Release Progress

Count
7,195,950
Goal is 50 Million

Count

IA files by OA status

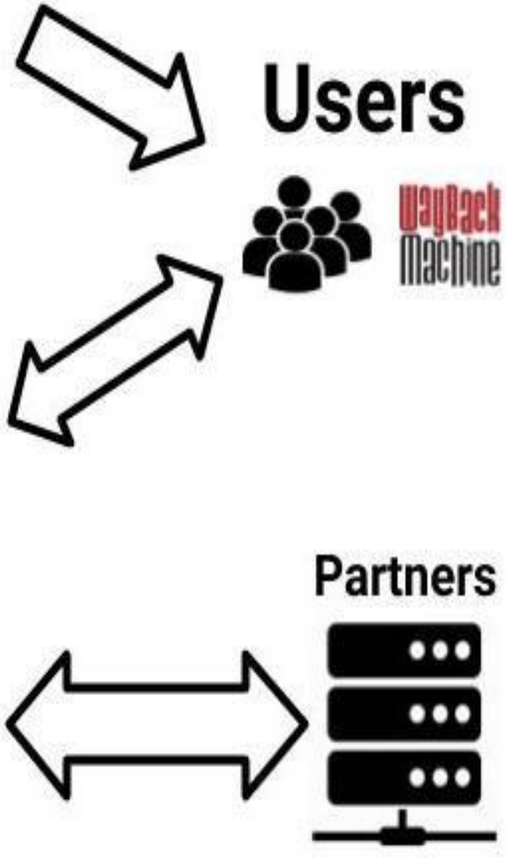
Long-Tail Fulltext Count

1,476,578
Long-Tail Releases With Fulltext

Release Count Total

81,357,478
Release Metadata Records (all types)

Release Types (with/without fulltext)

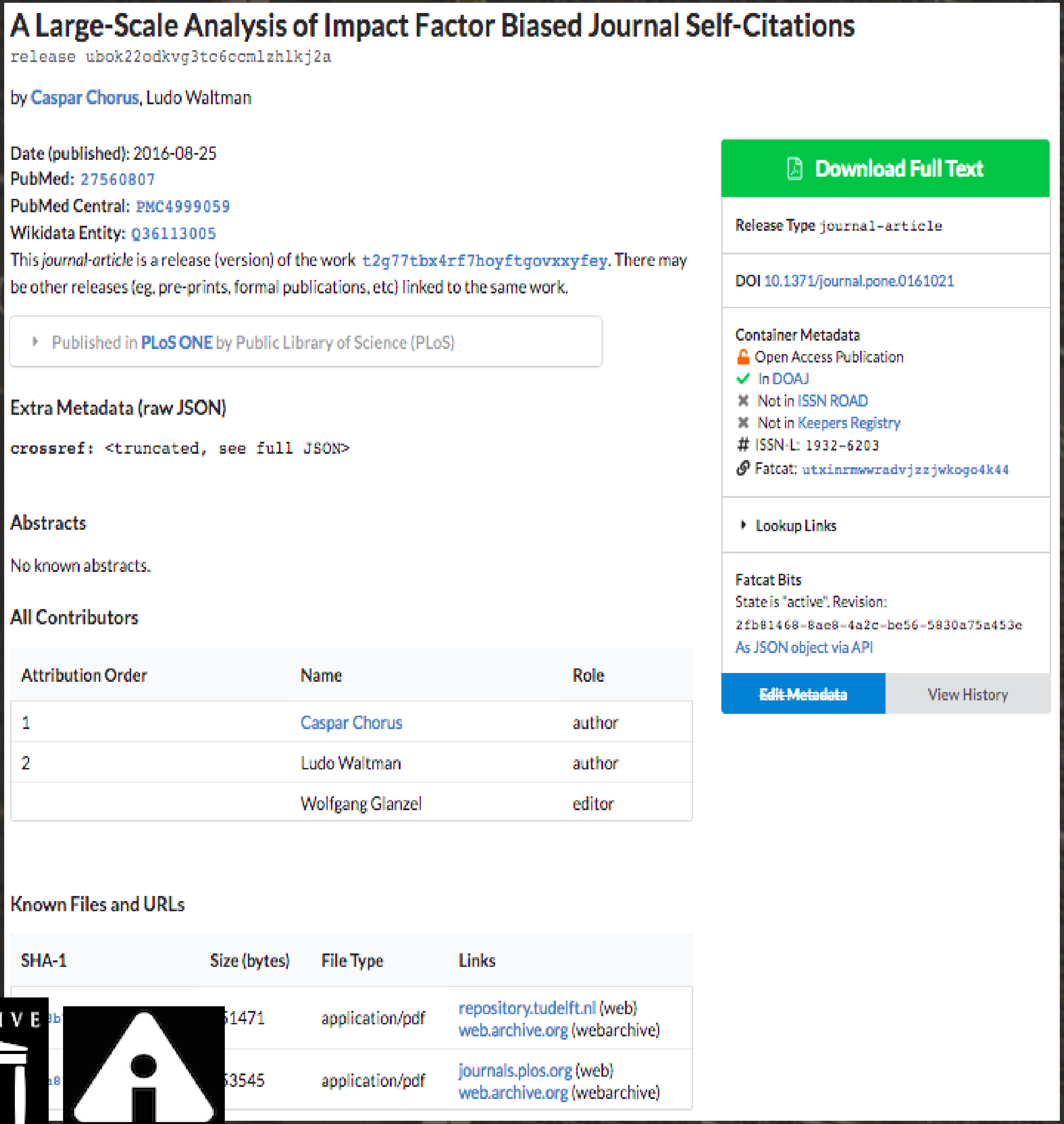


INTERNET ARCHIVE

ARCHIVE-IT

IIPC WAC 2019

Oh Look -- A GUI!



A Large-Scale Analysis of Impact Factor Biased Journal Self-Citations

`release ubok22odkvg3tc6ccmlzhlkj2a`

by Caspar Chorus, Ludo Waltman

Date (published): 2016-08-25
PubMed: `27560807`
PubMed Central: `PMC4999059`
Wikidata Entity: `Q36113005`
This *journal-article* is a release (version) of the work `t2g77tbx4rf7hoyftgovxxyfey`. There may be other releases (eg, 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

Attribution Order	Name	Role
1	Caspar Chorus	author
2	Ludo Waltman	author
	Wolfgang Glanzel	editor

Known Files and URLs

SHA-1	Size (bytes)	File Type	Links
3b	51471	application/pdf	repository.tudelft.nl (web) web.archive.org (webarchive)
a8	53545	application/pdf	journals.plos.org (web) web.archive.org (webarchive)

Download Full Text

Release Type `journal-article`

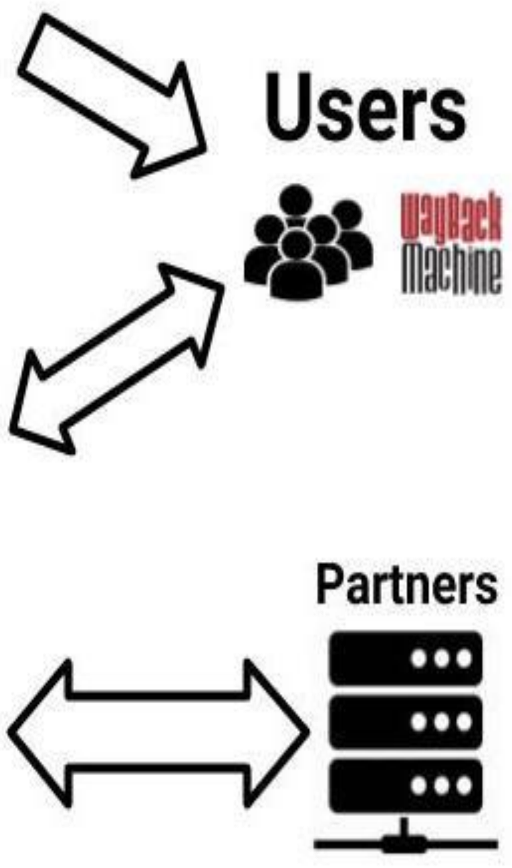
DOI 10.1371/journal.pone.0161021

Container Metadata
🔓 Open Access Publication
✓ In DOAJ
✖ Not in ISSN ROAD
✖ Not in Keepers Registry
ISSN-L: `1932-6203`
🔗 Fatcat: `utxinrmwwradvjzzjwkogo4k44`

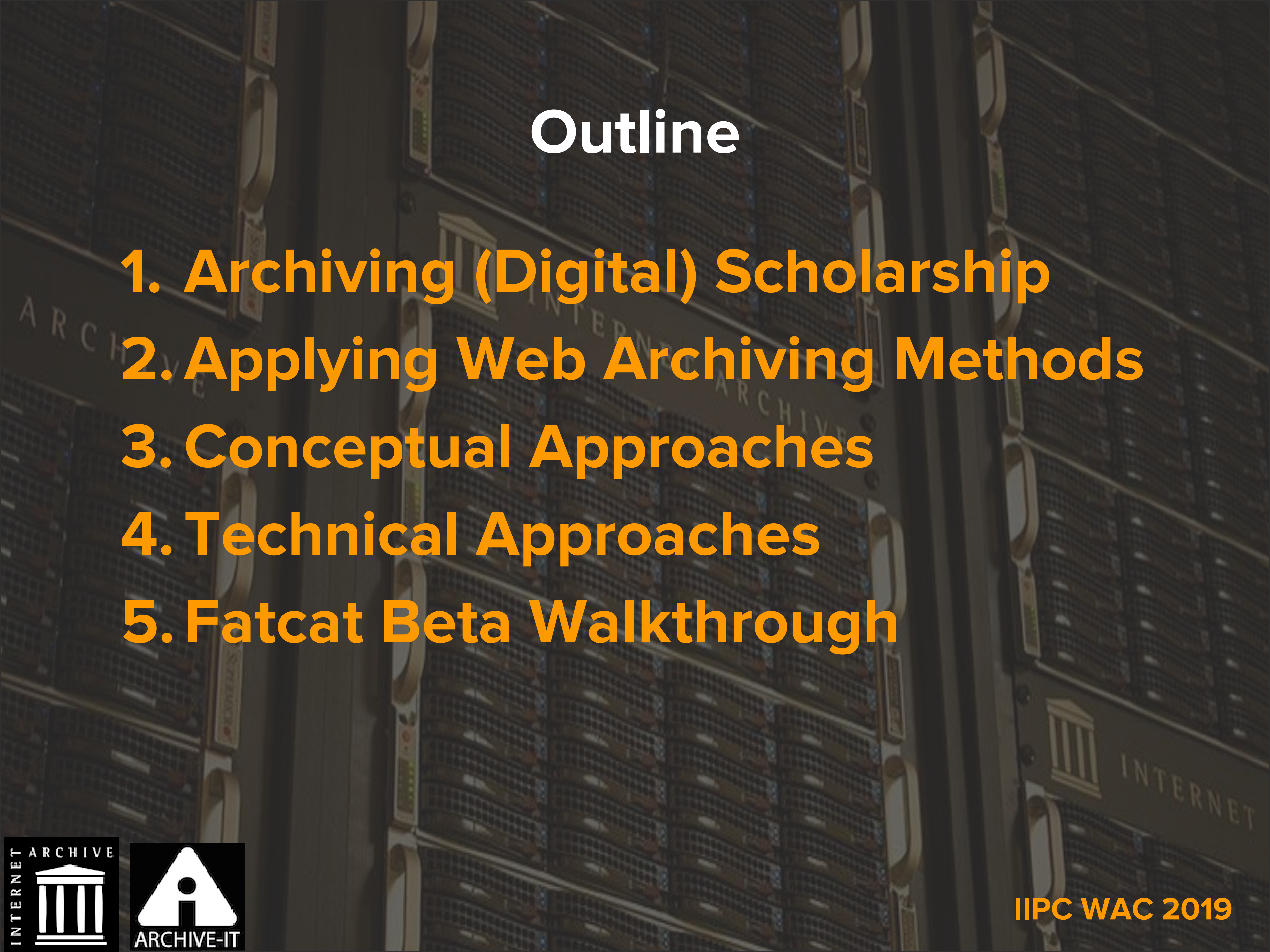
▸ Lookup Links

Fatcat Bits
State is "active". Revision:
`2fb81468-8ae8-4a2c-be56-5830a75a453e`
As JSON object via API

Edit Metadata | View History



IIPC WAC 2019

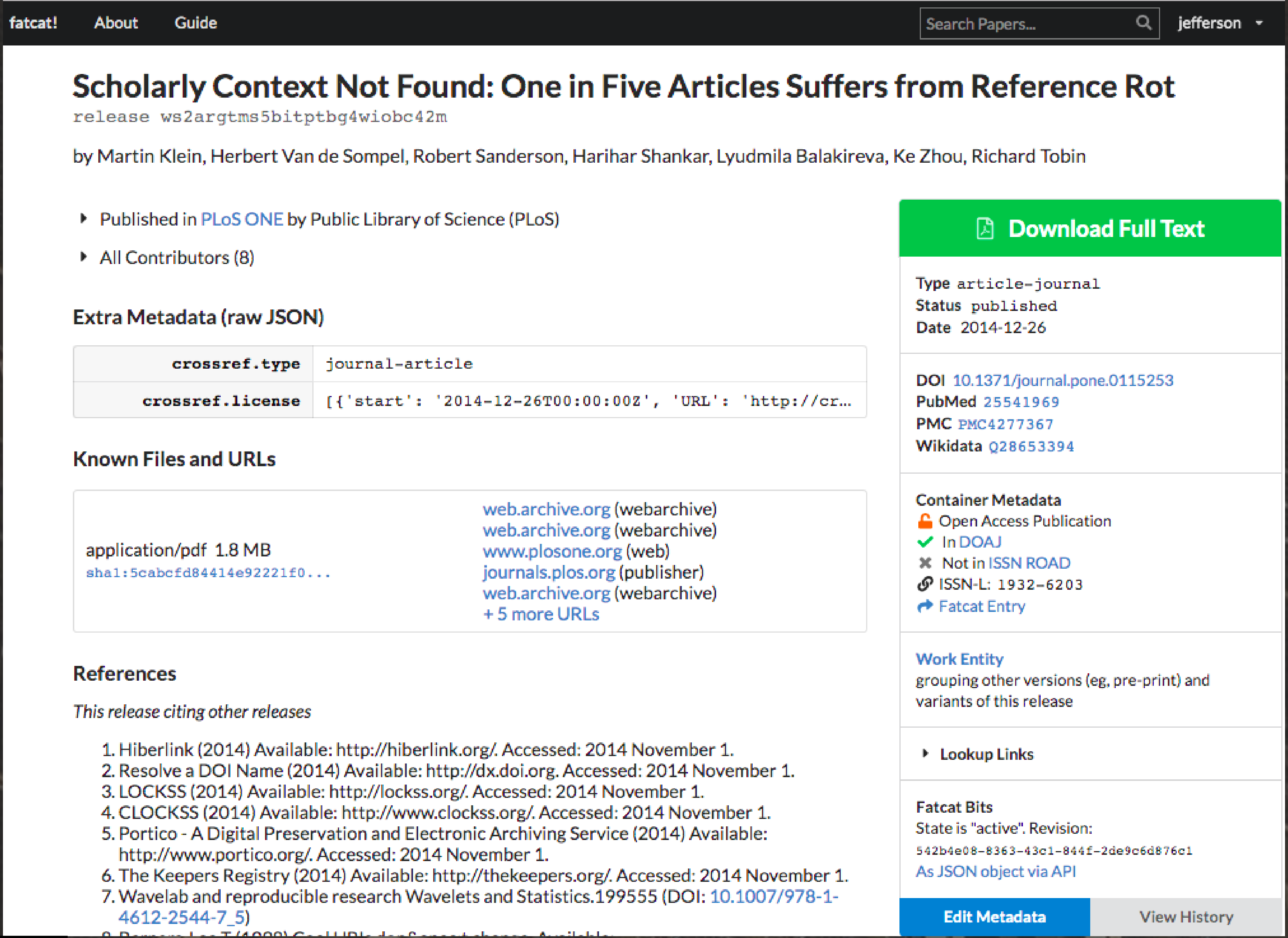


Outline

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

release ws2argtms5bitptbg4wiobc42m

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 sha1:5cabcfd84414e92221f0...	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

1. Hiberlink (2014) Available: http://hiberlink.org/. Accessed: 2014 November 1.
2. Resolve a DOI Name (2014) Available: http://dx.doi.org. Accessed: 2014 November 1.
3. LOCKSS (2014) Available: http://lockss.org/. Accessed: 2014 November 1.
4. CLOCKSS (2014) Available: http://www.clockss.org/. Accessed: 2014 November 1.
5. Portico - A Digital Preservation and Electronic Archiving Service (2014) Available: http://www.portico.org/. Accessed: 2014 November 1.
6. The Keepers Registry (2014) Available: http://thekeepers.org/. Accessed: 2014 November 1.
7. Wavelab and reproducible research Wavelets and Statistics.199555 (DOI: 10.1007/978-1-4612-2544-7_5)

Type article-journal
Status published
Date 2014-12-26

DOI 10.1371/journal.pone.0115253
PubMed 25541969
PMC PMC4277367
Wikidata Q28653394

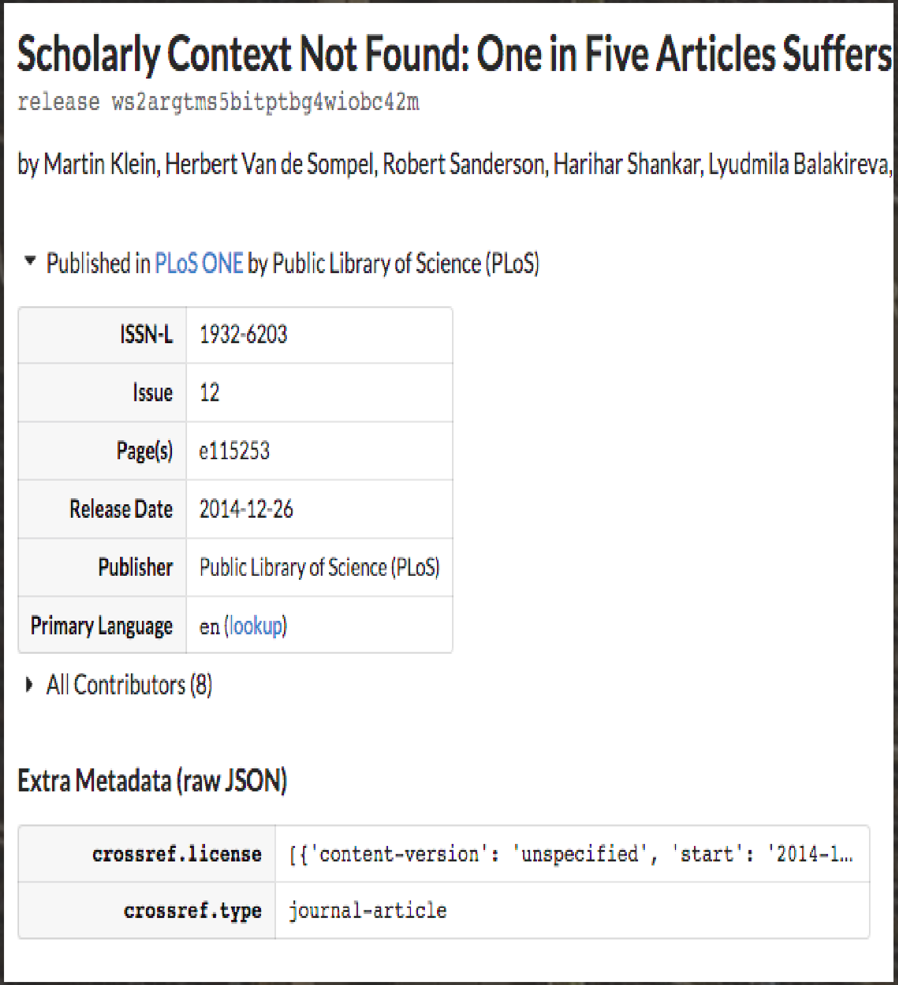
Container Metadata
Open Access Publication
✓ In DOAJ
✗ Not in ISSN ROAD
ISSN-L: 1932-6203
Fatcat Entry

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

Fatcat Bits
State is "active". Revision:
542b4e08-8363-43c1-844f-2de9c6d876c1
As JSON object via API

https://fatcat.wiki/

IIPC WAC 2019



Scholarly Context Not Found: One in Five Articles Suffers

release ws2argtms5bitptbg4wiobc42m

by Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva,

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

ISSN-L	1932-6203
Issue	12
Page(s)	e115253
Release Date	2014-12-26
Publisher	Public Library of Science (PLoS)
Primary Language	en (lookup)

▸ All Contributors (8)

Extra Metadata (raw JSON)

crossref.license	[{'content-version': 'unspecified', 'start': '2014-1...
crossref.type	journal-article

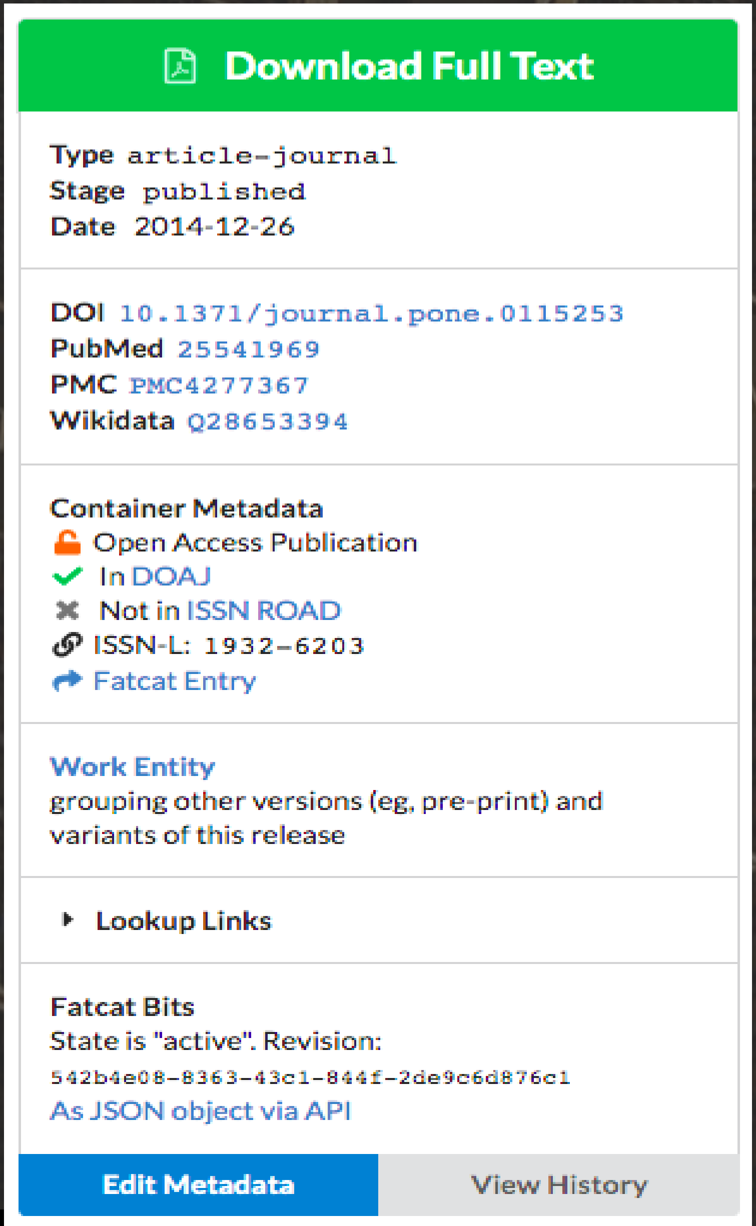
Bibliographic metadata

ISSN metadata

DOI metadata

https://fatcat.wiki/

IIPC WAC 2019



Download Full Text

Type article-journal
Stage published
Date 2014-12-26

DOI 10.1371/journal.pone.0115253
PubMed 25541969
PMC PMC4277367
Wikidata Q28653394

Container Metadata
🔓 Open Access Publication
✓ In DOAJ
✖ Not in ISSN ROAD
🔗 ISSN-L: 1932-6203
➦ Fatcat Entry

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

▸ Lookup Links

Fatcat Bits
State is "active". Revision:
542b4e08-8363-43c1-844f-2de9c6d876c1
As JSON object via API

Edit Metadata | View History

Download full text

Identifier linking

Registries lookup

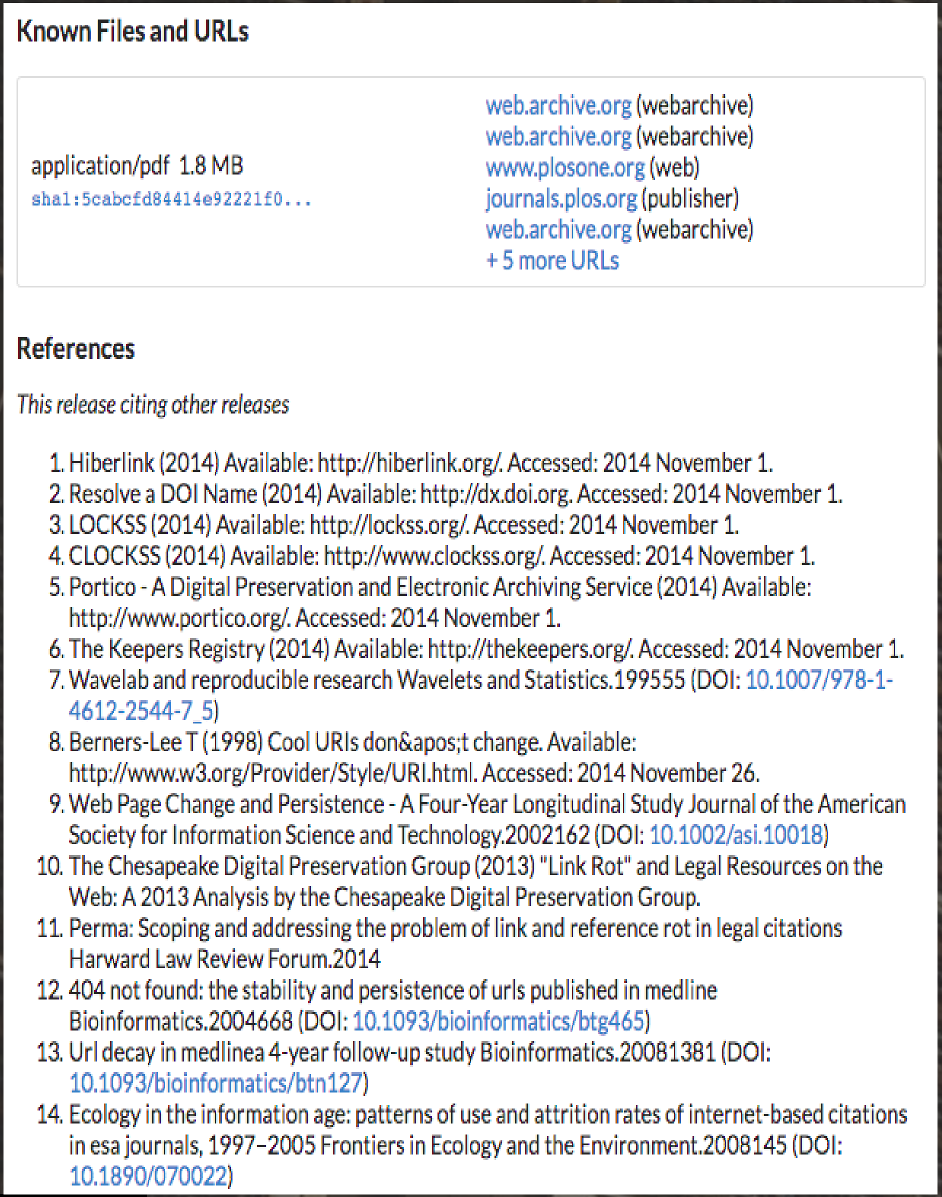
Version linking

JSON API

Record Editing

https://fatcat.wiki/

IIPC WAC 2019



Known Files and URLs

application/pdf 1.8 MB
sha1:5cabcfd84414e92221f0...

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

1. Hiberlink (2014) Available: http://hiberlink.org/. Accessed: 2014 November 1.
2. Resolve a DOI Name (2014) Available: http://dx.doi.org. Accessed: 2014 November 1.
3. LOCKSS (2014) Available: http://lockss.org/. Accessed: 2014 November 1.
4. CLOCKSS (2014) Available: http://www.clockss.org/. Accessed: 2014 November 1.
5. Portico - A Digital Preservation and Electronic Archiving Service (2014) Available: http://www.portico.org/. Accessed: 2014 November 1.
6. The Keepers Registry (2014) Available: http://thekeepers.org/. Accessed: 2014 November 1.
7. Wavelab and reproducible research Wavelets and Statistics.199555 (DOI: 10.1007/978-1-4612-2544-7_5)
8. Berners-Lee T (1998) Cool URIs don&apos;t change. Available: http://www.w3.org/Provider/Style/URI.html. Accessed: 2014 November 26.
9. Web Page Change and Persistence - A Four-Year Longitudinal Study Journal of the American Society for Information Science and Technology.2002162 (DOI: 10.1002/asi.10018)
10. The Chesapeake Digital Preservation Group (2013) "Link Rot" and Legal Resources on the Web: A 2013 Analysis by the Chesapeake Digital Preservation Group.
11. Perma: Scoping and addressing the problem of link and reference rot in legal citations Harward Law Review Forum.2014
12. 404 not found: the stability and persistence of urls published in medline Bioinformatics.2004668 (DOI: 10.1093/bioinformatics/btg465)
13. Url decay in medlinea 4-year follow-up study Bioinformatics.20081381 (DOI: 10.1093/bioinformatics/btn127)
14. Ecology in the information age: patterns of use and attrition rates of internet-based citations in esa journals, 1997–2005 Frontiers in Ecology and the Environment.2008145 (DOI: 10.1890/070022)

Wayback(!) and live web URLs + mime, size, checksum

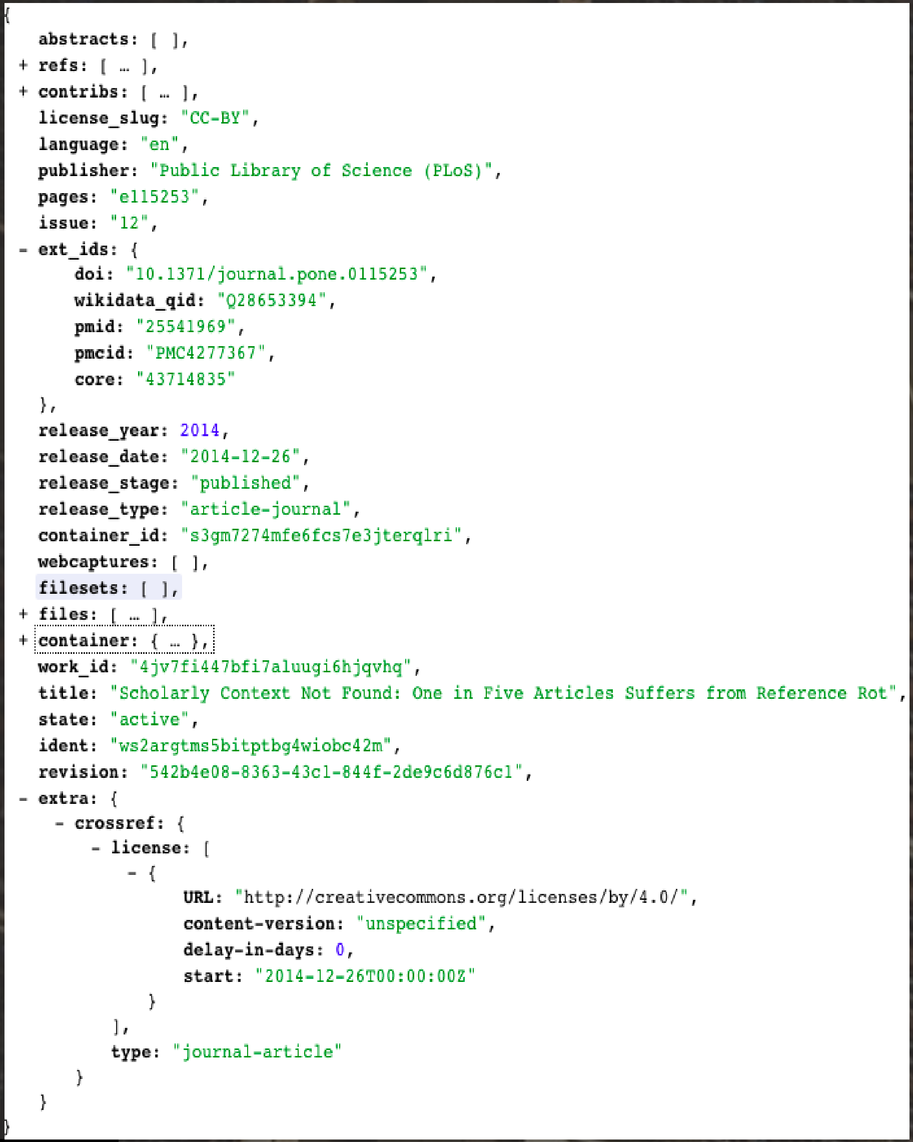
Extracted citations (interlinked to other fatcat records and wayback URLs for web references)

INTERNET ARCHIVE

ARCHIVE-IT

https://fatcat.wiki/

IIPC WAC 2019

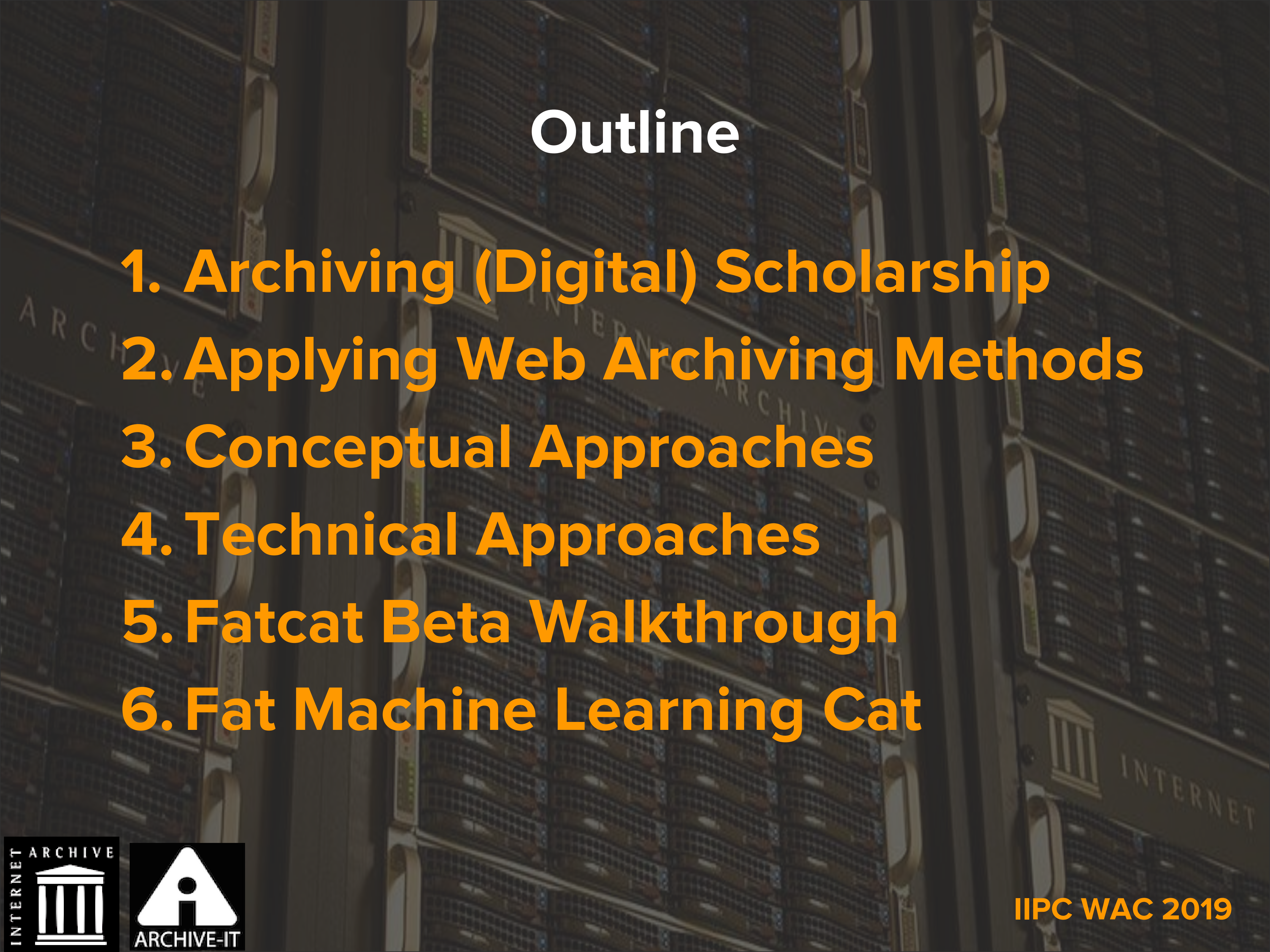


```
{
   abstracts: [ ],
 + refs: [ … ],
 + contribs: [ … ],
   license_slug: "CC-BY",
   language: "en",
   publisher: "Public Library of Science (PLoS)",
   pages: "e115253",
   issue: "12",
 - ext_ids: {
       doi: "10.1371/journal.pone.0115253",
       wikidata_qid: "Q28653394",
       pmid: "25541969",
       pmcid: "PMC4277367",
       core: "43714835"
   },
   release_year: 2014,
   release_date: "2014-12-26",
   release_stage: "published",
   release_type: "article-journal",
   container_id: "s3gm7274mfe6fcs7e3jterqlri",
   webcaptures: [ ],
   filesets: [ ],
 + files: [ … ],
 + container: { … },
   work_id: "4jv7fi447bfi7aluugi6hjqvhq",
   title: "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot",
   state: "active",
   ident: "ws2argtms5bitptbg4wiobc42m",
   revision: "542b4e08-8363-43c1-844f-2de9c6d876c1",
 - extra: {
     - crossref: {
         - license: [
             - {
                   URL: "http://creativecommons.org/licenses/by/4.0/",
                   content-version: "unspecified",
                   delay-in-days: 0,
                   start: "2014-12-26T00:00:00Z"
               }
           ],
           type: "journal-article"
       }
   }
}
```

The API, which has additional metadata not in the user interface

https://fatcat.wiki/

IIPC WAC 2019



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 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 substes to partners for local use/preservation
- Computational research services

IIPC WAC 2019

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
  - https://blog.dshr.org/2019/04/personal-pods-and-fatcat.html
- 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