Pathways to Preservation:

Article Processing Workflows for the CLOCKSS Archive

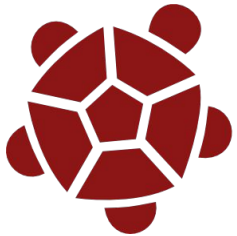
Meghan Frazer
Operations Manager
frazerm@stanford.edu

Carmen Cox
Plugin Developer
crc10@stanford.edu

Mary-Ellen Petrich
Digital Preservation Analyst
mpetrich@stanford.edu

What is LOCKSS?

- Based at Stanford University Libraries
- "Lots Of Copies Keep Stuff Safe"
- Provides a range of technical and operational support services to preservation networks powered by LOCKSS
- Develops the open source LOCKSS software
  - distributed digital preservation system
  - Highly resistant polling-and-repair protocol from peer-reviewed research
  - Web crawler, polling-and-repair preservation repository, metadata extraction, Web replay; driven by a plugin system



LOCKSS

What's with the turtle?

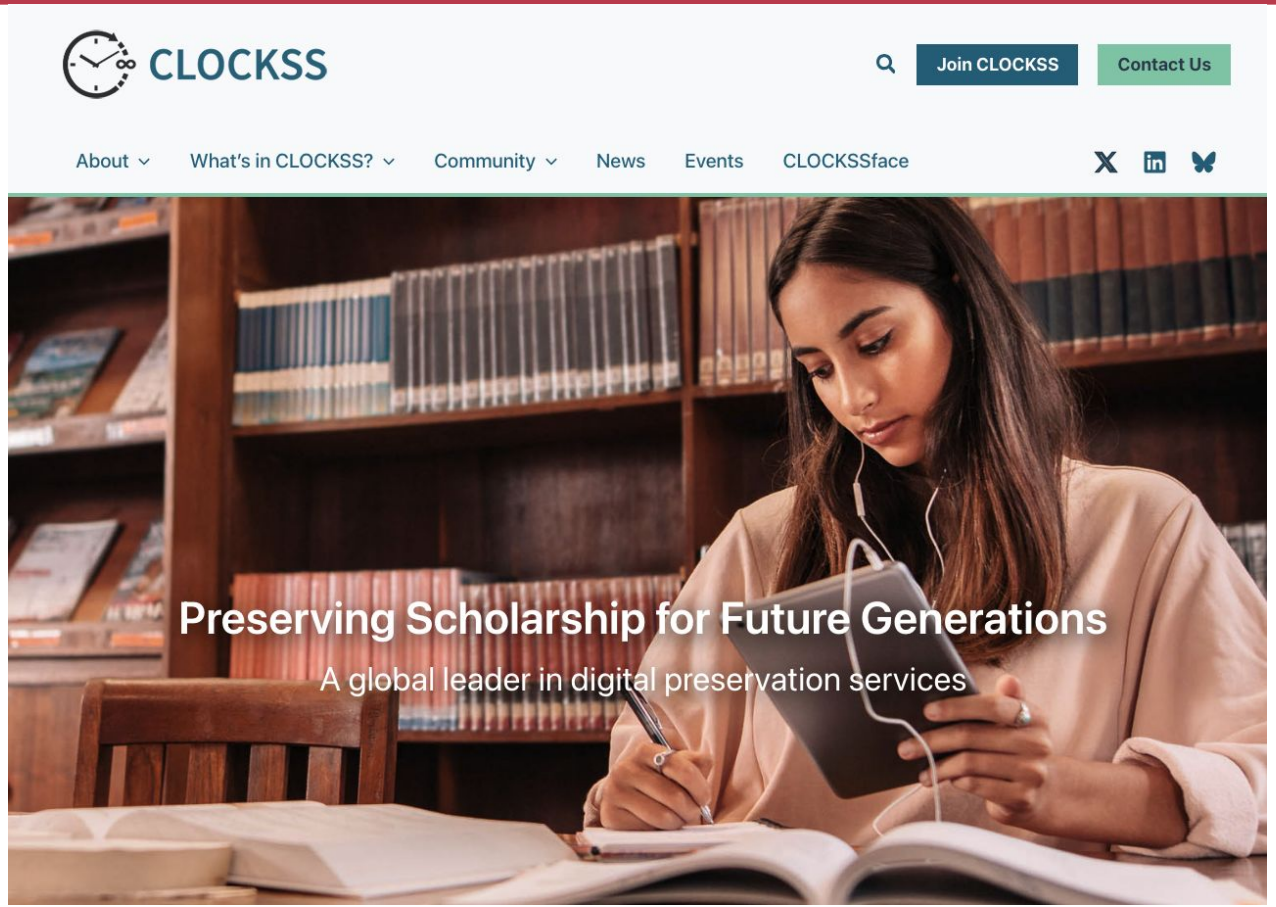
The LOCKSS logo reflects our core principles of community, decentralization, and durability. You may see it as:

A tortoise — In folklore tortoises represent deliberateness, determination, and longevity

Peers seated around a table — Digital preservation is most sustainably situated in a community with clear lines of communication and a shared commitment to saving content

Nodes in a network — LOCKSS is digital preservation in which multiple systems are all working together

What is CLOCKSS?



What is CLOCKSS? 5 essential facts

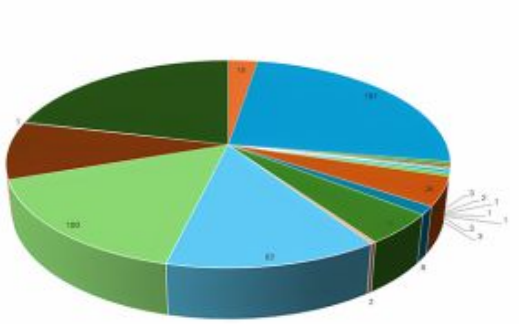
1. It's a dark archive.

2. It's international.

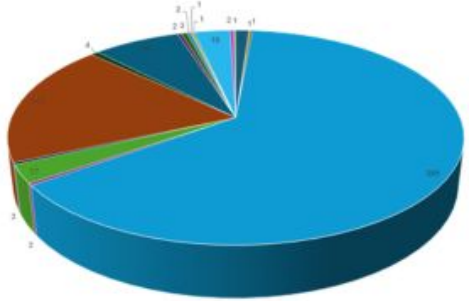
3. It's multi-disciplinary.

4. It's multilingual.

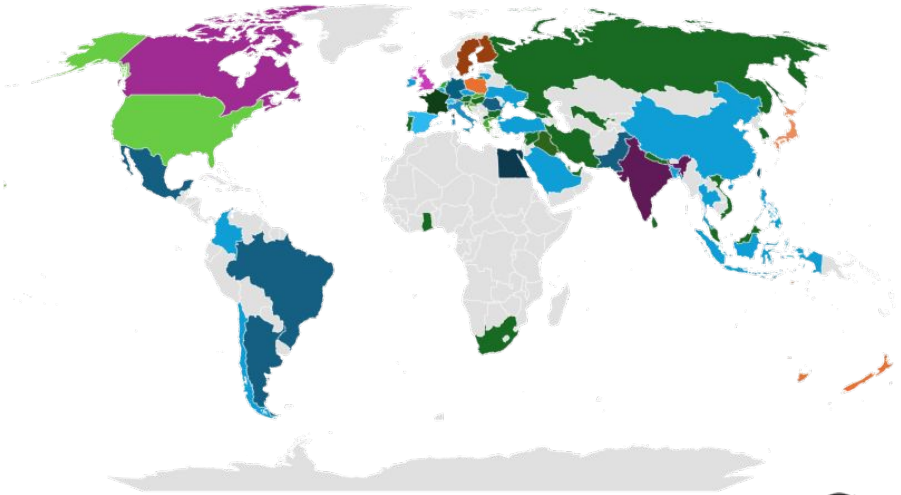
5. It's dedicated to diversity, equity and inclusion.



disciplines



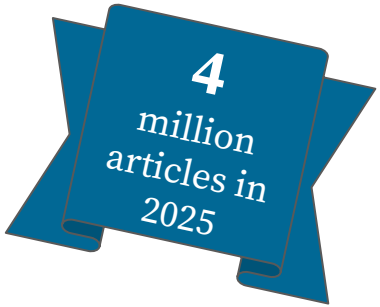
languages



Source: https://clockss.org/about/

What is CLOCKSS? By the Numbers

62.3 Million
Journal Articles



300
Supporting Libraries

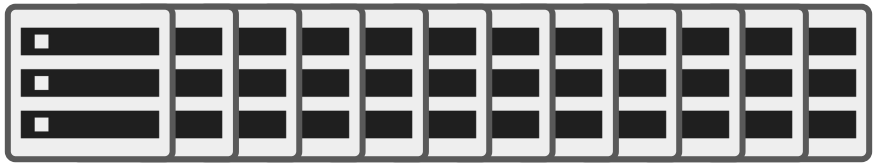
561,500
Books

681
Participating Publishers

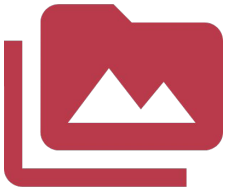
80
Triggered Titles Open Access

12
Mirror Repository Sites

~300
TB stored in each node and growing each year



Building Blocks: the AU



Archival Unit (AU):

A collection of files that can be crawled and has a defined start and endpoint.

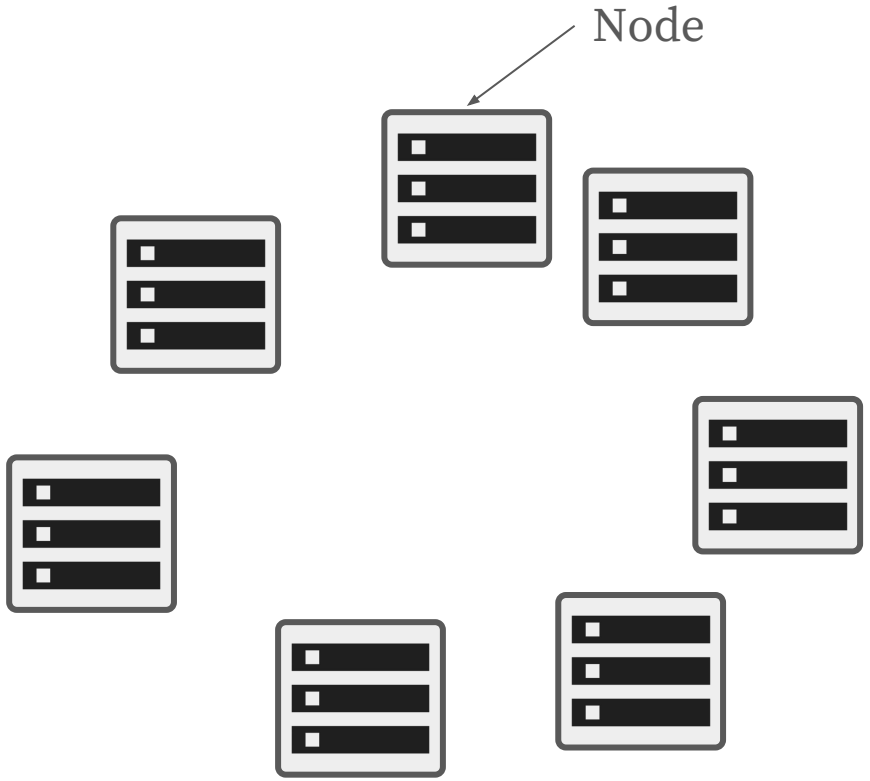
Examples:

- All content delivered from a single publisher via file transfer in 2026
- The 2025 volume of a single journal published on a publisher's website

Building Blocks: A LOCKSS Network

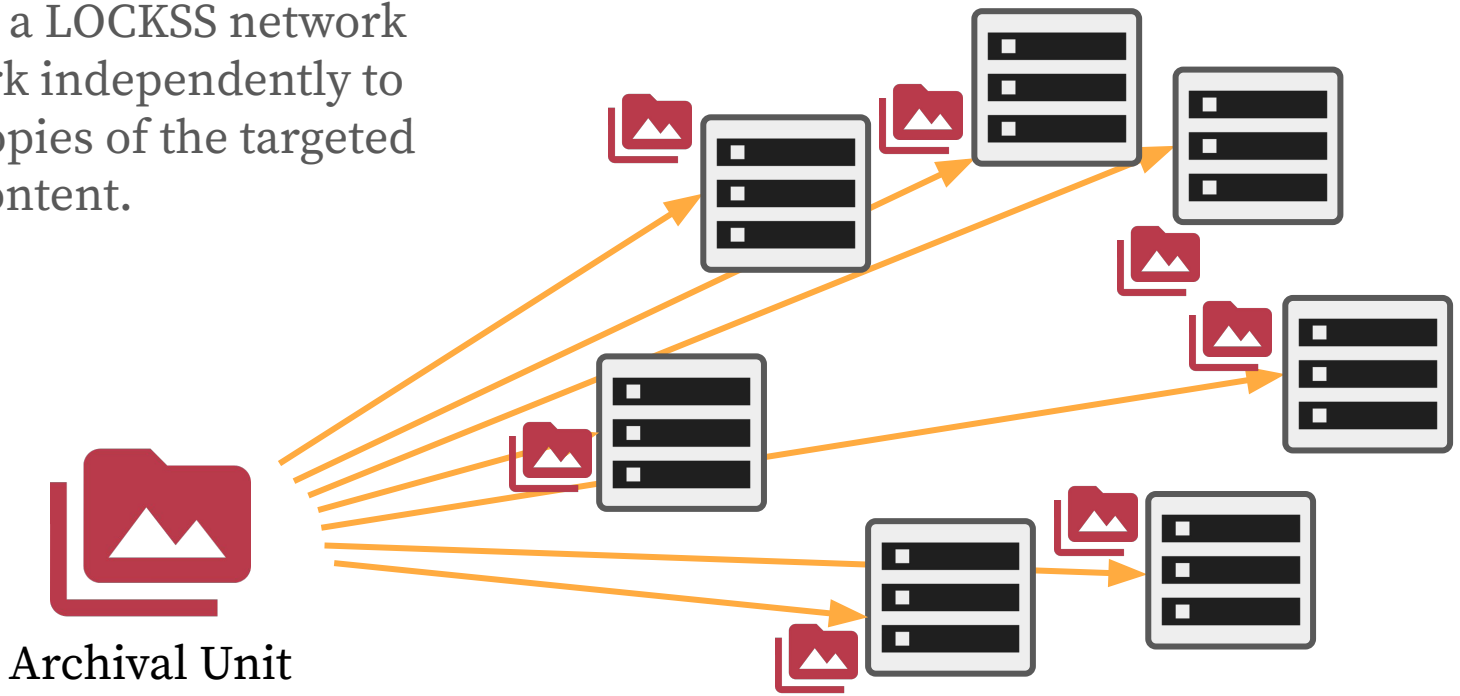
A small number of sites host hardware that runs the LOCKSS software and is dedicated for network use. This protects data through geographic separation and diversifies localized risks.

Archival Unit



Building Blocks: A LOCKSS Network

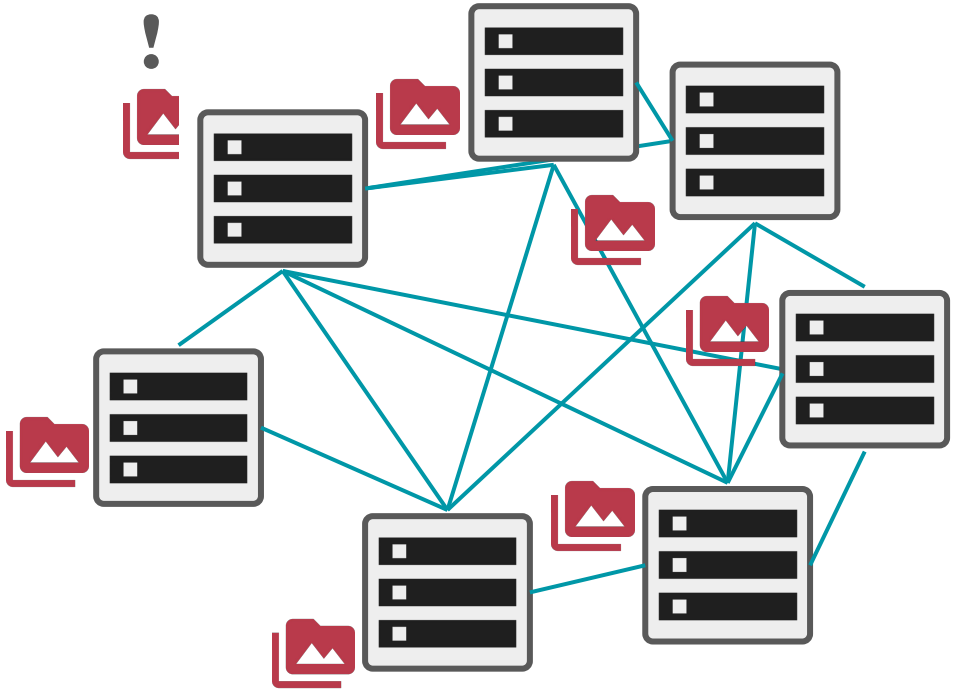
Nodes in a LOCKSS network each work independently to obtain copies of the targeted digital content.



Building Blocks: A LOCKSS Network

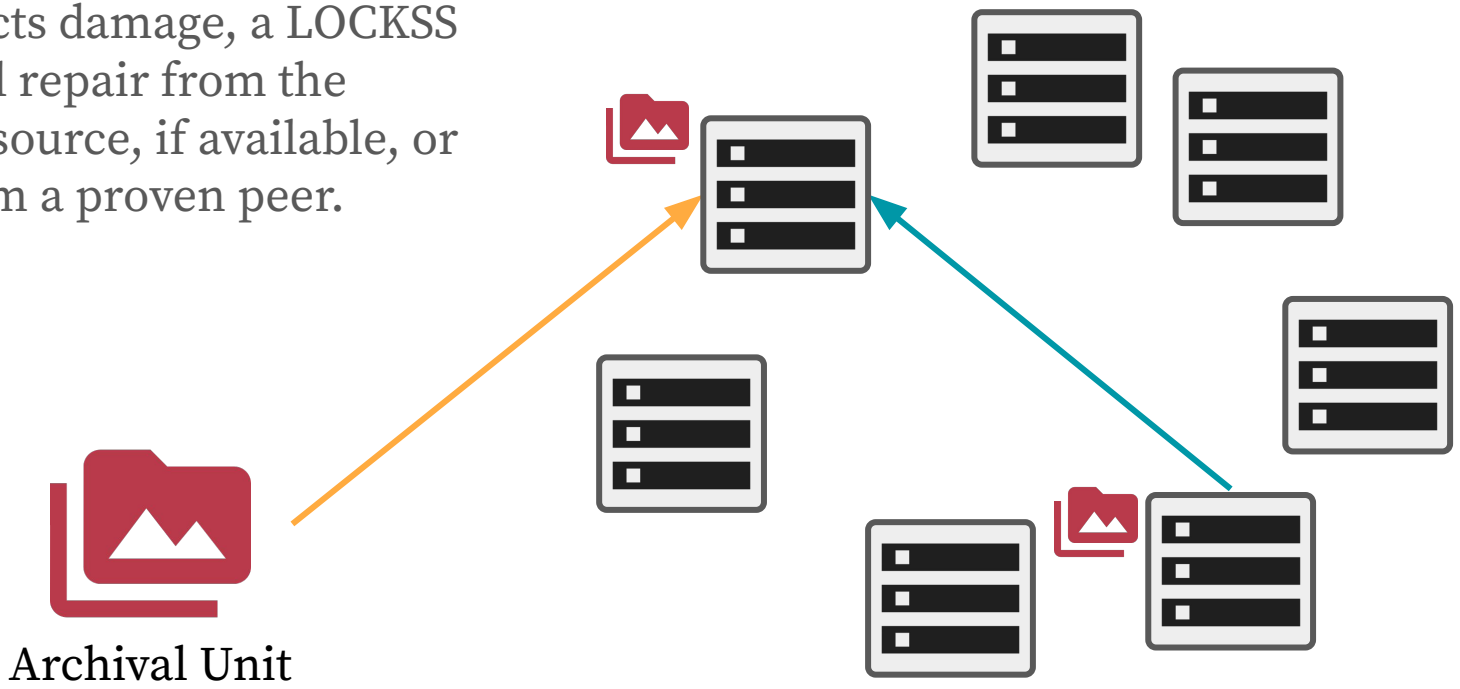
During the polling-and-repair process, nodes communicate with one another using a peer-to-peer protocol to verify that each stored copy is authentic and intact.

Archival Unit



Building Blocks: A LOCKSS Network

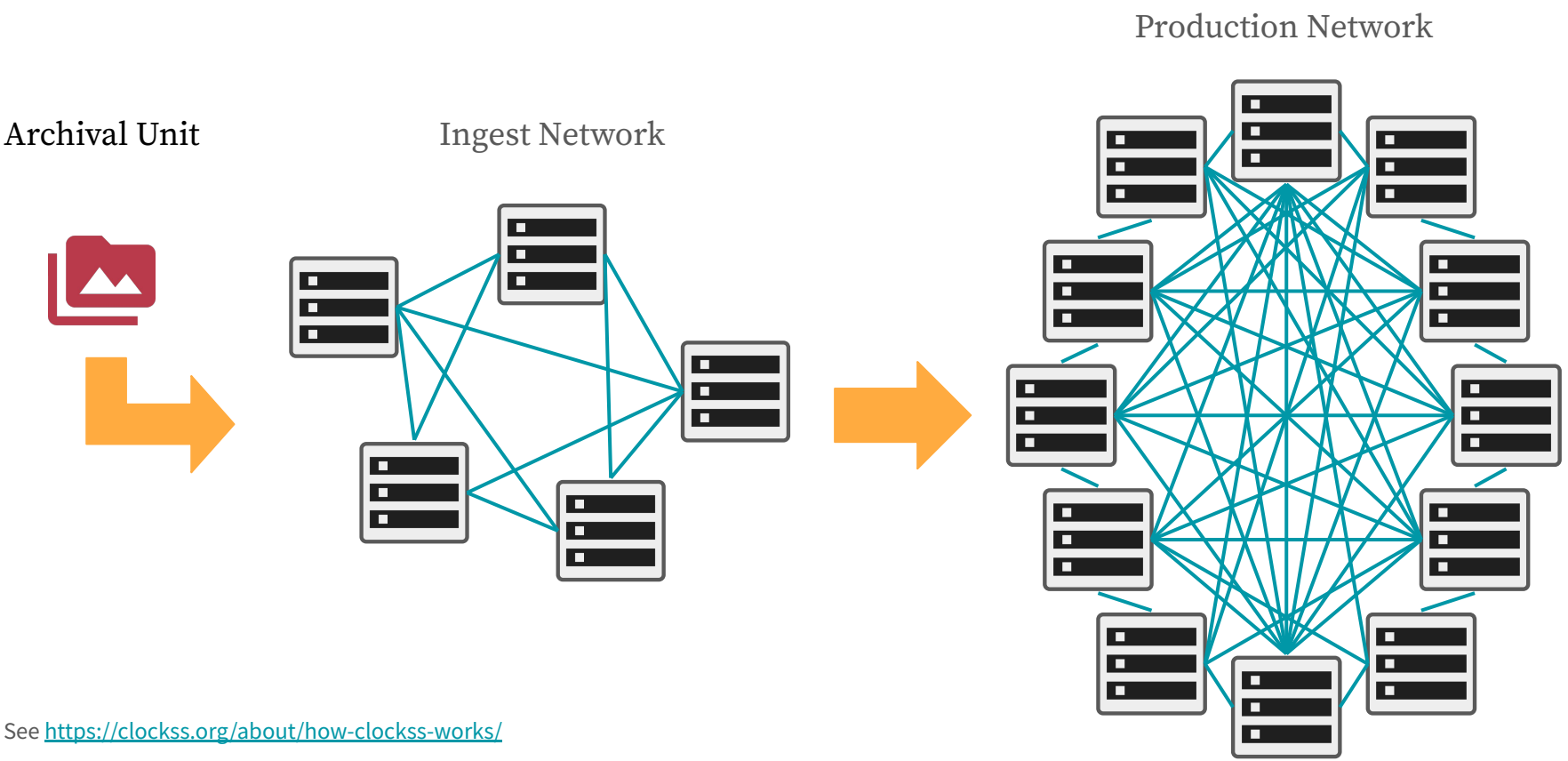
If it detects damage, a LOCKSS node will repair from the original source, if available, or copy from a proven peer.



A day in the life of a LOCKSS web crawler

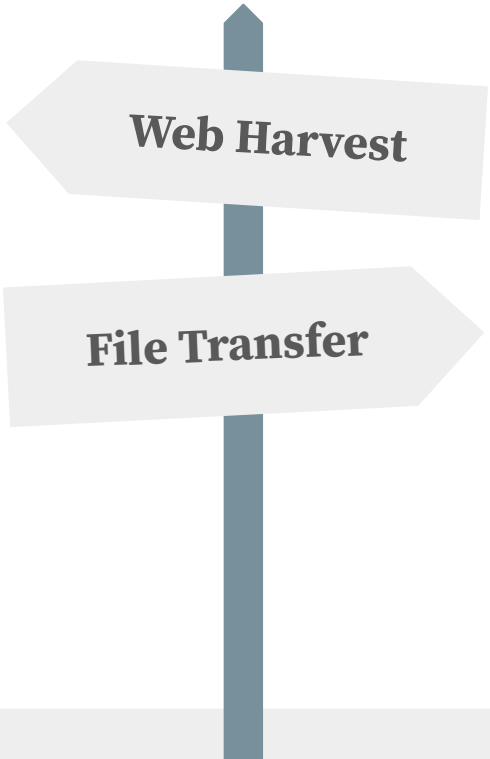
- Crawl
  - Documents are collected from the publisher website
- Poll
  - A quorum of machines compares successful crawls
- Hash
  - Compare to each other
- Repair
  - Disagreements are resolved with a repair from the original source if available

CLOCKSS infrastructure



See https://clockss.org/about/how-clockss-works/

Web Harvest vs File Transfer Pathways



Web Harvest vs File Transfer Pathways

AUs: ~202K in the Archive

- Volume of a Journal
- Volume or Chapter of a Book
- A finite collection of documents
- Up to ~500 GB

Access via IP subscription and the presence of the CLOCKSS Permission Statement on any domain with content or web replay files

Start page

- List of journal issues by year or volume
- Allows for discovery of a limited amount of content at a time.

Web Harvest vs File Transfer Pathways

AUs: ~1600

- Content delivered by partner
- Divided into buckets up to ~300 GB
- Large publishers may deliver many AUs per year
- > 80% of articles delivered this way

Delivery via FTP, SFTP, Snail Mail

Metadata standards and data structure are established during setup and need to stay consistent. All metadata comes from delivered files.

WARCs and Project Jasper are special types of File Transfer where the content is staged by us (WARC), or delivered in partnership with DOAJ and the Internet Archive (Project Jasper).

Web Harvest vs File Transfer Onboarding

What are the titles, URLs, and identifiers (ISBN or ISSN)?

Where is the metadata?

Is there more than one host (for images, scripts)?

Is there a lot of dynamic content?

How is the content divided? By issue, volume, or is it continuous?

Open-access or IP-based subscription access?

What are the titles, and identifiers (ISBN or ISSN)?

What content and metadata formats will be sent (PDFs, epubs, XML, JATS metadata, ONIX metadata, etc)?

How are the content and metadata related to each other in the delivery?

How frequently will content be delivered?

How much will content be delivered in a year?

Evaluation and Plugins

Evaluation & Plugin Development

What is a plugin?

- A LOCKSS plugin is a bundle of rules and code adapting the general LOCKSS software to the particulars of a given preservation target, such as a publishing platform (ex: Open Journal Systems)
- Java + xml configuration files

Evaluation & Plugin Development

What is a plugin?

- A LOCKSS plugin is a bundle of rules and code adapting the general LOCKSS software to the particulars of a given preservation target, such as a publishing platform (ex: Open Journal Systems)
- Java + xml configuration files

```
<entry>
  <string>plugin_version</string>
  <string>71</string>
</entry>
 <entry>
  <string>plugin_name</string>
  <string>Open Journal Systems Plugin (OJS 3.x)</string>
</entry>
<entry>
  <string>au_name</string>
  <string>"Open Journal Systems Plugin (OJS 3.x), Base URL %s, Journal ID %s, Year %d", base_url, journal_id, year</string>
</entry>
<entry>
  <string>plugin_config_props</string>
  <list>
     <org.lockss.daemon.ConfigParamDescr>
        <key>year</key>
        <displayName>Year</displayName>
        <description>Four digit year (e.g., 2004)</description>
        <type>4</type>
        <size>4</size>
        <definitional>true</definitional>
        <defaultOnly>false</defaultOnly>
     </org.lockss.daemon.ConfigParamDescr>
     <org.lockss.daemon.ConfigParamDescr>
        <key>journal_id</key>
        <displayName>Journal Identifier</displayName>
        <description>Identifier for journal (often used as part of file names)</description>
        <type>1</type>
        <size>40</size>
        <definitional>true</definitional>
        <defaultOnly>false</defaultOnly>
     </org.lockss.daemon.ConfigParamDescr>
     <org.lockss.daemon.ConfigParamDescr>
        <key>base_url</key>
        <displayName>Base URL</displayName>
        <description>Usually of the form http://&lt;journal-name&gt;.com/</description>
        <type>3</type>
        <size>40</size>
        <definitional>true</definitional>
        <defaultOnly>false</defaultOnly>
     </org.lockss.daemon.ConfigParamDescr>
  </list>
</entry>
```

Evaluation & Plugin Development

How do we get content?

- File Transfer (FTP)
- Web Harvest

Index of /sourcefiles/oup-released/2025_01

- abmv59i1.zip
- abmv59i1.zip.md5sum
- abtv8i4.zip
- abtv8i4.zip.md5sum
- adaptationv19i1.zip
- adaptationv19i1.zip.md5sum
- aev71i4.zip
- aev71i4.zip.md5sum
- ageingv54i10.zip
- ageingv54i10.zip.md5sum
- ageingv54i12.zip
- ageingv54i12.zip.md5sum
- ageingv54iSupplement_4.zip
- ageingv54iSupplement_4.zip.md5sum
- ajeadvancesv1i3.zip
- ajeadvancesv1i3.zip.md5sum
- ajhpv82i24.zip
- ajhpv82i24.zip.md5sum
- ajhpv83i1.zip
- ajhpv83i1.zip.md5sum
- alcalcv61i1.zip
- alcalcv61i1.zip.md5sum
- amtv50i1.zip
- amtv50i1.zip.md5sum
- annwehv69i9.zip
- annwehv69i9.zip.md5sum
- aobplav17i6.zip
- aobplav17i6.zip.md5sum
- aobv136i5-6.zip
- aobv136i5-6.zip.md5sum
- arthistoryv48i4.zip
- arthistoryv48i4.zip.md5sum
- asjopenforumv7.zip
- asjopenforumv7.zip.md5sum
- asjv45i11.zip
- asjv45i11.zip.md5sum
- bcsjv98i12.zip
- bcsjv98i12.zip.md5sum
- behecov36i6.zip
- behecov36i6.zip.md5sum
- behecov37i1.zip
- behecov37i1.zip.md5sum
- bfgv24.zip
- bfgv24.zip.md5sum
- bibv26i6.zip
- bibv26i6.zip.md5sum

Evaluation & Plugin Development

Your turn...

Supplementary Materials

The PDF file includes:

Supplementary Methods
Supplementary Discussion
Figs. S1 to S29
Table S3
References (*102*–*108*)

DOWNLOAD | 12.16 MB

Other Supplementary Material for this manuscript includes the following:

Tables S1 and S2

DOWNLOAD | 120.65 KB

MDAR Reproducibility Checklist

DOWNLOAD | 463.92 KB

REFERENCES AND NOTES

1. E. S. Gray, M. C. Madiga, T. Hermanus, P. L. Moore, C. K. Wibmer, N. L. Tumba, L. Werner, K. Mlisana, S. Sibeko, C. Williamson, S. S. A. Karim, L. Morris, CAPRISA002 Study Team, The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4$^+$ T cell decline and high viral load during acute infection. *J. Virol.* **85**, 4828–4840 (2011).

GO TO REFERENCE • CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR • Find it at Stanford

2. N. A. Doria-Rose, R. M. Klein, M. G. Daniels, S. O'Dell, M. Nason, A. Lapedes, T. Bhattacharya, S. A. Migueles, R. T. Wyatt, B. T. Korber, J. R. Mascola, M. Connors, Breadth of human immunodeficiency virus-specific neutralizing activity in sera: Clustering analysis and association with clinical variables. *J. Virol.* **84**, 1631–1636 (2010).

CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR • Find it at Stanford

RECOMMENDED

RESEARCH ARTICLE | JULY 2015
Priming a broadly neutralizing antibody response to HIV-1 using a germline-targeting immunogen

REPORT | MARCH 2016
HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen

REPORT | AUGUST 2010
Rational Design of Envelope Identifies Broadly Neutralizing Human Monoclonal Antibodies to HIV-1

RESEARCH ARTICLE | MARCH 2017
Mimicry of an HIV broadly neutralizing antibody epitope with a synthetic glycopeptide

SPONSORED WEBINAR | SCIENCE AND LIFE | 18 DEC 2025
Building a global community for rare disease: Accelerating treatments, access, and collaboration

Supplementary Materials

The PDF file includes:

Supplementary Methods
Supplementary Discussion
Figs. S1 to S29
Table S3
References (*102*–*108*)

DOWNLOAD 12.16 MB

Other Supplementary Material for this manuscript includes the following:

Tables S1 and S2

DOWNLOAD 120.65 KB

MDAR Reproducibility Checklist

DOWNLOAD 463.92 KB

We want to collect these.

We don't want to collect these.

REFERENCES AND NOTES

1. E. S. Gray, M. C. Madiga, T. Hermanus, P. L. Moore, C. K. Wibmer, N. L. Tumba, L. Werner, K. Mlisana, S. Sibeko, C. Williamson, S. S. A. Karim, L. Morris, CAPRISA002 Study Team, The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4$^+$ T cell decline and high viral load during acute infection. *J. Virol.* **85**, 4828–4840 (2011).

GO TO REFERENCE • CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR • Find it at Stanford

2. N. A. Doria-Rose, R. M. Klein, M. G. Daniels, S. O'Dell, M. Nason, A. Lapedes, T. Bhattacharya, S. A. Migueles, R. T. Wyatt, B. T. Korber, J. R. Mascola, M. Connors, Breadth of human immunodeficiency virus-specific neutralizing activity in sera: Clustering analysis and association with clinical variables. *J. Virol.* **84**, 1631–1636 (2010).

CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR • Find it at Stanford

Evaluation & Plugin Development

Evaluation

- How dynamic is the website?

Dynamic	Static
Generates content on the fly. Generally more difficult to fully capture all elements of a website.	Delivers pre-built HTML files "as-is". Generally easier for the crawler to fully capture.

Evaluation & Plugin Development

Evaluation

- Is there metadata? How is it formatted?

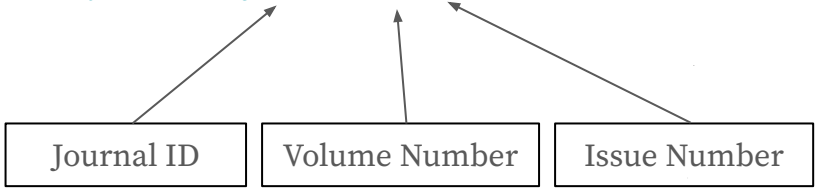
```
<meta name="citation_journal_title" content="Transactions on Energy Systems and Engineering Applications"/>
<meta name="citation_journal_abbrev" content="Trans. Energy Syst. Eng. Appl."/>
<meta name="citation_issn" content="2745-0120"/>
<meta name="citation_author" content="Oscar Acevedo"/>
<meta name="citation_author_institution" content="Universidad Tecnológica de Bolívar"/>
<meta name="citation_title" content="Engineering the Future: TESEA's Commitment to Quality and Innovation"/>
<meta name="citation_language" content="en"/>
<meta name="citation_date" content="2024/06/30"/>
<meta name="citation_volume" content="5"/>
<meta name="citation_issue" content="1"/>
<meta name="citation_firstpage" content="1"/>
<meta name="citation_lastpage" content="1"/>
<meta name="citation_doi" content="10.32397/tesea.vol5.n1.705"/>
<meta name="citation_abstract_html_url" content="https://revistas.utb.edu.co/tesea/article/view/705"/>
<meta name="citation_keywords" xml:lang="en" content="Indexing"/>
<meta name="citation_keywords" xml:lang="en" content="Journal rank"/>
<meta name="citation_keywords" xml:lang="en" content="academic publishing"/>
<meta name="citation_pdf_url" content="https://revistas.utb.edu.co/tesea/article/download/705/398"/>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE article PUBLIC "-//NLM//DTD JATS (Z39.96) Journal Publishing DTD v1.0 20120330//EN"
<article article-type="research-article" xml:lang="en"
    xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <?origin annotum?>
    <front>
        <journal-meta>
            <journal-id journal-id-type="publisher"/>
            <journal-title-group>
                <journal-title>Open Health Data</journal-title>
            </journal-title-group>
            <issn>2054-7102</issn>
            <publisher>
                <publisher-name>Ubiquity Press</publisher-name>
            </publisher>
        </journal-meta>
        <article-meta>
            <article-id pub-id-type="doi">10.5334/ohd.ag</article-id>
            <article-categories>
                <subj-group>
                    <subject>Data paper</subject>
                </subj-group>
            </article-categories>
            <title-group>
                <article-title>A First Attempt at Modelling Red Deer (<italic>Cervus
                    elaphus)</italic> Distributions Over Europe</article-title>
            </title-group>
            <contrib-group>
                <contrib>
                    <name>
                        <surname>Wint</surname>
                        <given-names>William</given-names>
                    </name>
                    <xref ref-type="aff" rid="aff-1"/>
                </contrib>
```

Evaluation & Plugin Development

Evaluation

- Is the URL structure consistent and predictable?
  - Predictable URL: https://ascelibrary.org/toc/jbenf2/22/12



Evaluation & Plugin Development

What does a complete plugin do?

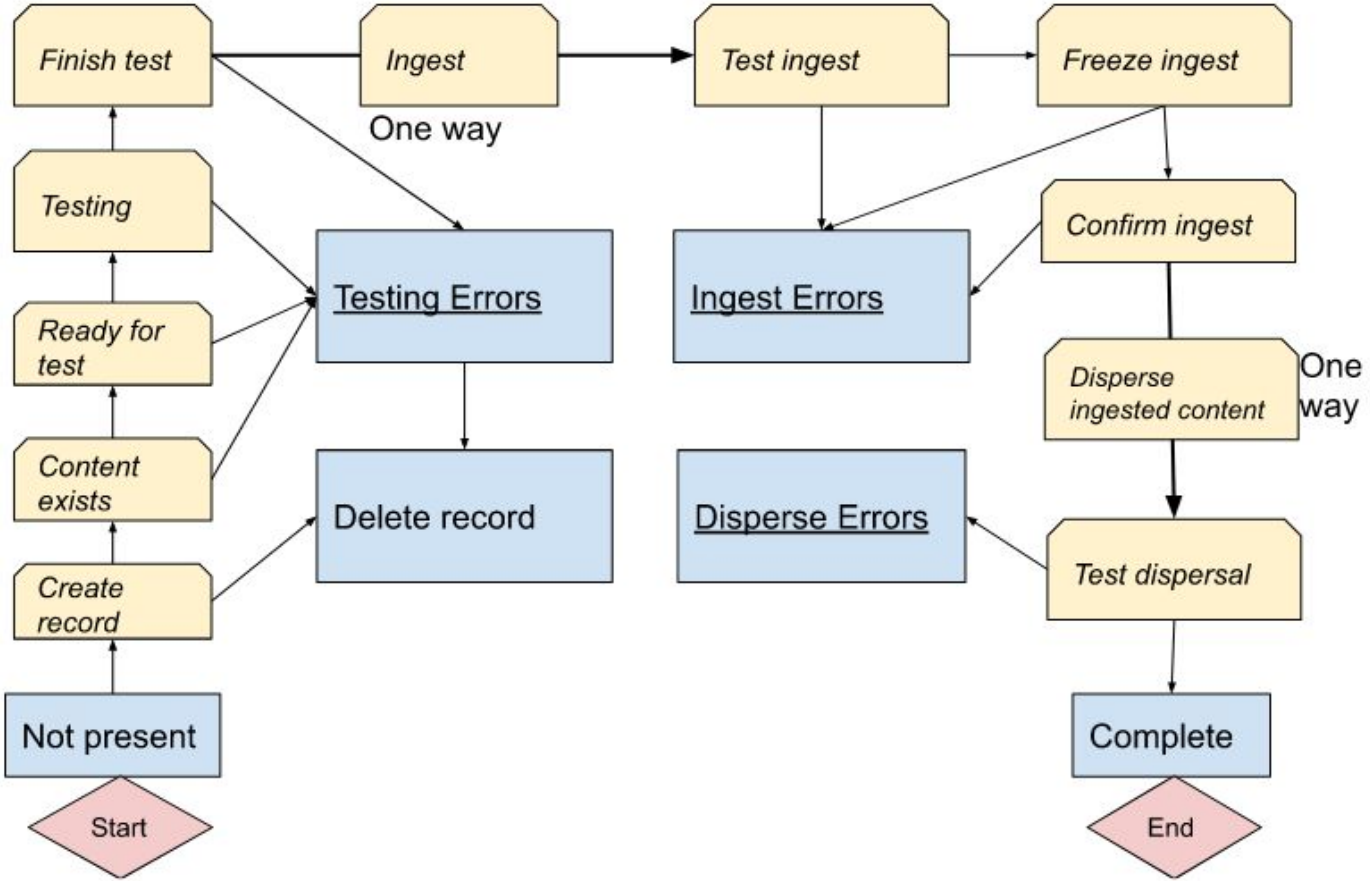
- Collects all scholarly content specific to an Archival Unit
- Lists accurate number of articles
- Collects as much metadata as is available
- Provides representative replay of website
- Does not undercrawl or overcrawl

Evaluation & Plugin Development

- 70% of plugin development is maintenance
- Websites change over time

Content Test and Release

Content Testing & Content Release



Content Testing & Content Release

- ***expected*** not known if AU exists on the publisher's website
- ***exists*** known that AU exists on the publisher's web site
- ***manifest*** permission page and manifest verified
- ***testing*** a LOCKSS team member is testing this AU
- ***notReady*** testing has failed
- ***ready*** testing is complete and AU is ready for release to staging servers

- ***crawling*** iteratively crawling on the staging servers
- ***ingNotReady*** error found on the staging servers
- ***finished*** testing complete on staging servers
- ***down*** no longer crawling, awaiting resolution
- ***superseded*** this AU was found to be malformed

Content Testing

- Discover
  - Find content
- Document
  - Create specialized catalog records
- Test
  - Test crawls of content to ensure software compatibility
- Debug
- Release to the staging network
  - Continuously releasing year round, ~1000 archival units per month

Discover New Material

- add new publishers & journals (on demand)
- new volumes to add, predictable & unpredictable (annually)
- find new start pages (script checks 1x/wk)

Discover Maintenance Tasks

- old volumes have moved, developed errors
- merge metadata for multiple networks
- compare the catalog to the network
- QA. typos, duplicate ISSNs, duplicate volumes, malformed parameters

Document

- title database (tdb file)
- basic metadata
  - publisher, title, publication year, issn/isbn
  - in case metadata is missing from web site
- parameters for each AU
  - url & volume or year or others
  - defines the AUid
    - passes parameter values to the publisher plugin
    - unique key
- status
  - human readable
  - LOCKSS archive: recognize, crawl, don't crawl

```
{
  publisher <
    name = Taylor & Francis ;
    info[tester] = 6
  >

  plugin = org.lockss.plugin.taylorandfrancis.TaylorAndFrancisPlugin
  param[base_url] = http://www.tandfonline.com/

  {

    title <
      name = Archives and Manuscripts ;
      issn = 0157-6895 ;
      eissn = 2164-6058
    >

    param[journal_id] = raam20

    implicit < status ; year ; name ; param[volume_name] >

    au < manifest ; 2012 ; Archives and Manuscripts Volume 40 ; 40 >
    au < down ; 2013 ; Archives and Manuscripts Volume 41 ; 41 >
    au < down ; 2014 ; Archives and Manuscripts Volume 42 ; 42 >
    au < down ; 2015 ; Archives and Manuscripts Volume 43 ; 43 >
    au < manifest ; 2016 ; Archives and Manuscripts Volume 44 ; 44 >

  }
}
```

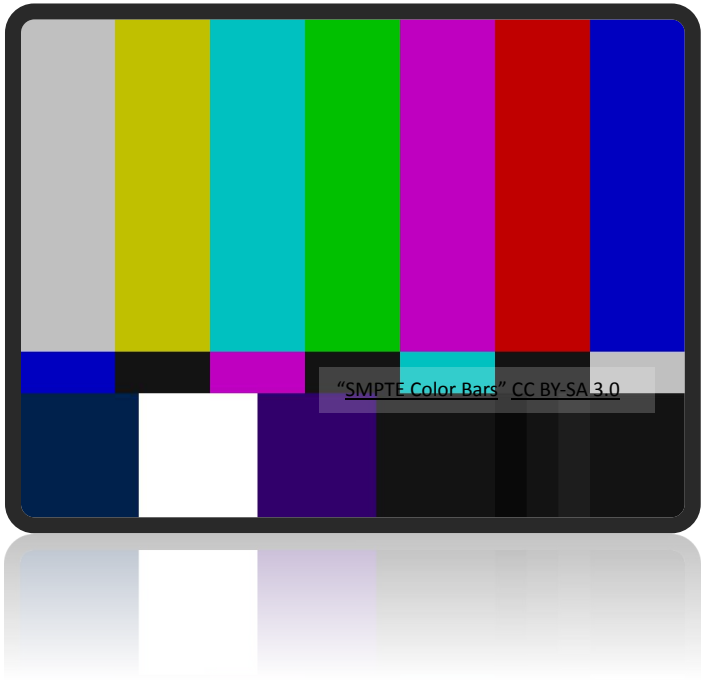
- Text file
- Human readable and editable
- Java scripts transform this data for:
  - searching & reports (csv)
  - LOCKSS archive (xml)
- shell, purl, python, awk scripts transform data


```
{
  publisher <
    name = Taylor & Francis ;
    info[taster] = 6
  >

  plugin = org.lockss.plugin.taylorandfrancis.TaylorAndFrancisPlugin
  param[base_url] = http://www.tandfonline.com/

  {

    title <
      name = Archives and Manuscripts ;
      issn = 0157-6895 ;
      eissn = 2164-6058
    >

    param[journal_id] = raam20

    implicit < status ; year ; name ; param[volume_name] >

    au < manifest ; 2012 ; Archives and Manuscripts Volume 40 ; 40 >
    au < down ; 2013 ; Archives and Manuscripts Volume 41 ; 41 >
    au < down ; 2014 ; Archives and Manuscripts Volume 42 ; 42 >
    au < down ; 2015 ; Archives and Manuscripts Volume 43 ; 43 >
    au < manifest ; 2016 ; Archives and Manuscripts Volume 44 ; 44 >

  }
}
```

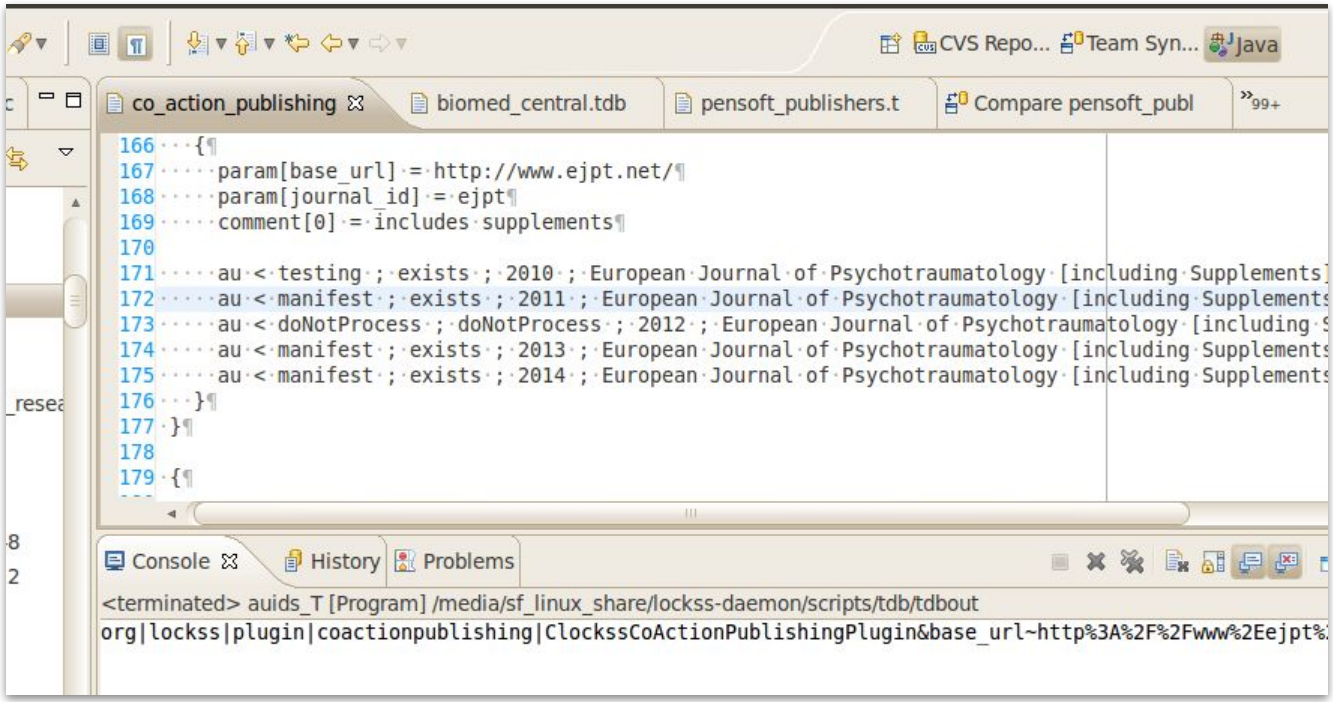
- Text file
- Human readable and editable
- Java scripts transform this data for:
  - searching & reports (csv)
  - LOCKSS archive (xml)
- shell, purl, python, awk scripts transform data further for complex reports

Testing

- Test content against software
  - two servers (choosing from 4)
  - 12 hours apart
- Errors
  - No subscription
  - Permission statement missing
  - No volume
  - Malformed lists of issues, articles, or links
  - URL redirects (journal has moved)
  - No articles
  - HTML crawl errors (can't access, taking too long, missing, moved)
  - Transient changes, rotating ads, dynamic content, watermarking



Testing: Generating an AUid



AUid

Add AUs

```
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~December&year~2021
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~January&year~2020
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~March&year~2020
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~April&year~2020
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~August&year~2020
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~December&year~2020
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~July&year~2020
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~May&year~2020
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~September&year~202
&MonthlyPlugin&base_url~https%3A%2F%2Fjournals%2Eplos%2Eorg%2F&journal_id~plosone&month~October&year~2020
```

☑ Start Tests ☐ Specify Daemons [ ] Keyword

Add AUs

Search Tests: 11 results found

AUID: PLoSMonthlyPlugin Keyword: Test State: Hashing / Finished / Error Created By: feili / crc10 / mellen Network: L / C Search ☐ Output Results as AUID List

☐	Network	AUTest State	Daemon Status	AU Name	Started	Created By	Notes
☐	CLOCKSS	Finished	99% (119426/119428) sullockss-content-03.stanford.edu:8081 [1285] sullockss-content-01.stanford.edu:8081 [1285]	PLOS ONE Volume Dec 2021	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (120878/120878) sullockss-content-02.stanford.edu:8081 [1323] sullockss-content-03.stanford.edu:8081 [1323]	PLOS ONE Volume Jan 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (112646/112646) sullockss-content-04.stanford.edu:8081 [1275] sullockss-content-02.stanford.edu:8081 [1275]	PLOS ONE Volume Mar 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (140804/140804) sullockss-content-01.stanford.edu:8081 [1558] sullockss-content-04.stanford.edu:8081 [1558]	PLOS ONE Volume Apr 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (132044/132044) sullockss-content-03.stanford.edu:8081 [1445] sullockss-content-01.stanford.edu:8081 [1445]	PLOS ONE Volume Aug 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (154422/154422) sullockss-content-04.stanford.edu:8081 [1735] sullockss-content-02.stanford.edu:8081 [1735]	PLOS ONE Volume Dec 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (119296/119296) sullockss-content-02.stanford.edu:8081 [1333] sullockss-content-03.stanford.edu:8081 [1333]	PLOS ONE Volume Jul 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (127540/127540) sullockss-content-01.stanford.edu:8081 [1415] sullockss-content-04.stanford.edu:8081 [1415]	PLOS ONE Volume May 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (137864/137864) sullockss-content-02.stanford.edu:8081 [1521] sullockss-content-01.stanford.edu:8081 [1521]	PLOS ONE Volume Sep 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (135658/135658) sullockss-content-03.stanford.edu:8081 [1501] sullockss-content-02.stanford.edu:8081 [1501]	PLOS ONE Volume Oct 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS
☐	CLOCKSS	Finished	100% (136580/136580) sullockss-content-04.stanford.edu:8081 [1503] sullockss-content-03.stanford.edu:8081 [1503]	PLOS ONE Volume Nov 2020	2026-01-08 13:44 Thu	Mary-Ellen	c_PLoS

Archive Selected Tests | Save Search | Recrawl | ReHealthCheck | Rehash

Successful Result

Crawl Tracker : Finished [-]
Crawl Tracker Created : Jan. 8, 2026, 1:58 p.m.
Crawl Tracker Finished : Feb. 6, 2026, 12:04 p.m.

Crawl on **sullockss-content-02.stanford.edu:8081** : Finished
Link to **Metadata**
Status on Daemon : Successful
AUTest Crawl Requested : Jan. 21, 2026, 1:24 p.m.
Daemon Crawl Started : Jan. 21, 2026, 1:23 p.m.
Daemon Crawl Finished : Jan. 25, 2026, 7:31 a.m.
Error, Fetched, Parsed and Excluded URL lists
Info/Error Message : AU has 1323 mdata, 1323 article, 1323 suburls, 60440 wurls, 60439 urls

Crawl on **sullockss-content-03.stanford.edu:8081** : Finished
Link to **Metadata**
Status on Daemon : Successful
AUTest Crawl Requested : Feb. 2, 2026, 3:30 p.m.
Daemon Crawl Started : Feb. 2, 2026, 3:30 p.m.
Daemon Crawl Finished : Feb. 6, 2026, 11:15 a.m.
Error, Fetched, Parsed and Excluded URL lists
Info/Error Message : AU has 1323 mdata, 1323 article, 1323 suburls, 60440 wurls, 60439 urls

Select Recrawl Daemons : [+]
Recrawl DELETE

Hash Tracker : Finished [+]

Hash Compare
100% (120878/120878)

AU-Test by LOCKSS

Unsuccessful Result

Crawl Tracker : Finished [-]
Crawl Tracker Created : Jan. 8, 2026, 1:58 p.m.
Crawl Tracker Finished : Feb. 3, 2026, 11:41 a.m.

Crawl on **sullockss-content-03.stanford.edu:8081** : Finished
Link to **Metadata**
Status on Daemon : Successful
AUTest Crawl Requested : Jan. 21, 2026, 10:54 a.m.
Daemon Crawl Started : Jan. 21, 2026, 10:54 a.m.
Daemon Crawl Finished : Jan. 25, 2026, 4 a.m.
Error, Fetched, Parsed and Excluded URL lists
Info/Error Message : AU has 1285 mdata, 1285 article, 1285 suburls, 59715 wurls, 59714 urls

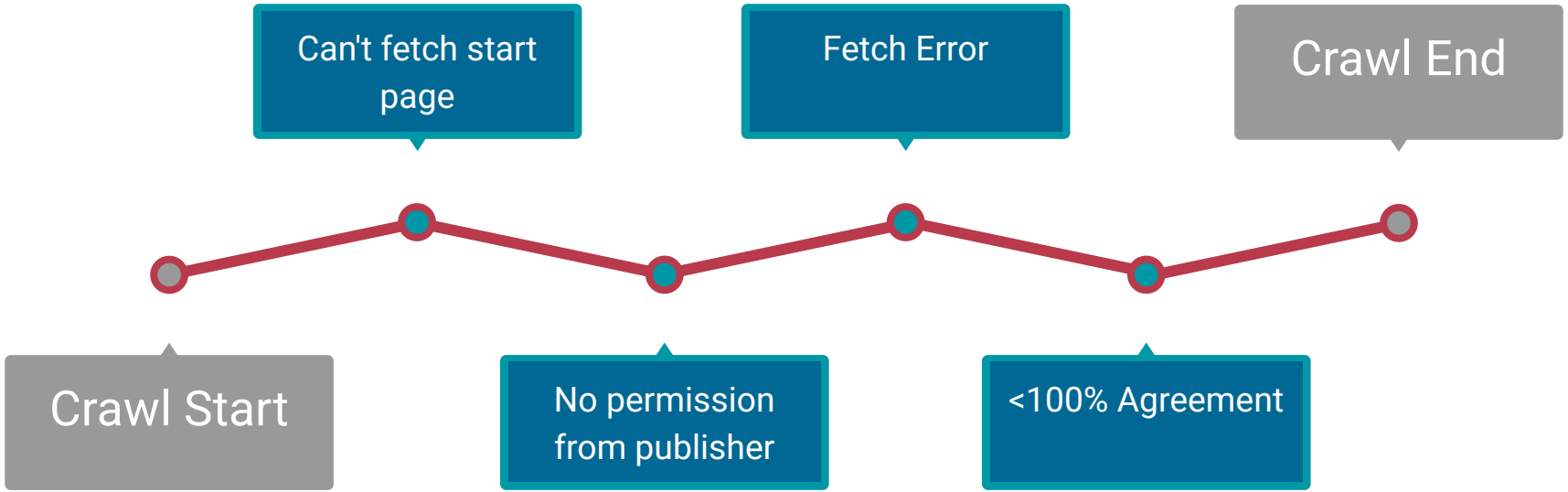
Crawl on **sullockss-content-01.stanford.edu:8081** : Finished
Link to **Metadata**
Status on Daemon : Successful
AUTest Crawl Requested : Jan. 30, 2026, 5:40 p.m.
Daemon Crawl Started : Jan. 30, 2026, 5:39 p.m.
Daemon Crawl Finished : Feb. 3, 2026, 10:40 a.m.
Error, Fetched, Parsed and Excluded URL lists
Info/Error Message : AU has 1285 mdata, 1285 article, 1285 suburls, 59715 wurls, 59714 urls

Select Recrawl Daemons : [+]
Recrawl DELETE

Hash Tracker : Finished [+]

Hash Compare
99% (119426/119428)
< > https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0261036 (text) (clearer text)
2 Different_urls

Categories of Errors



Debugging: Can't fetch start page

- Volume does not exist
  - Change status to doesNotExist.
- AUid formed incorrectly
  - Correct tdb file.
- Permission statement or start page missing
  - Request from publisher.
- Start page url redirects
  - Journal may have moved.

Debugging: No publisher permission

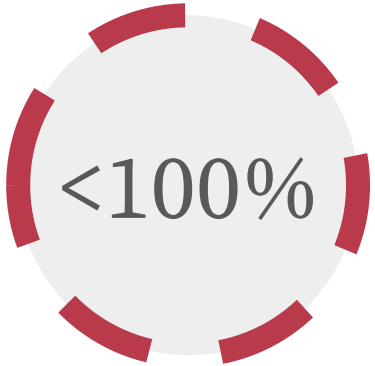
- No articles
  - No content to preserve; no subscription; or article URLs have changed.
- Publisher has delayed the release of content
  - Confirm that newer content is not publicly accessible.
- Crawler finds a login page on articles
  - Request subscription from publisher.
- Crawler succeeds on some test servers, but not on others
  - Request all LOCKSS IP addresses for subscription.

Debugging: Fetch error

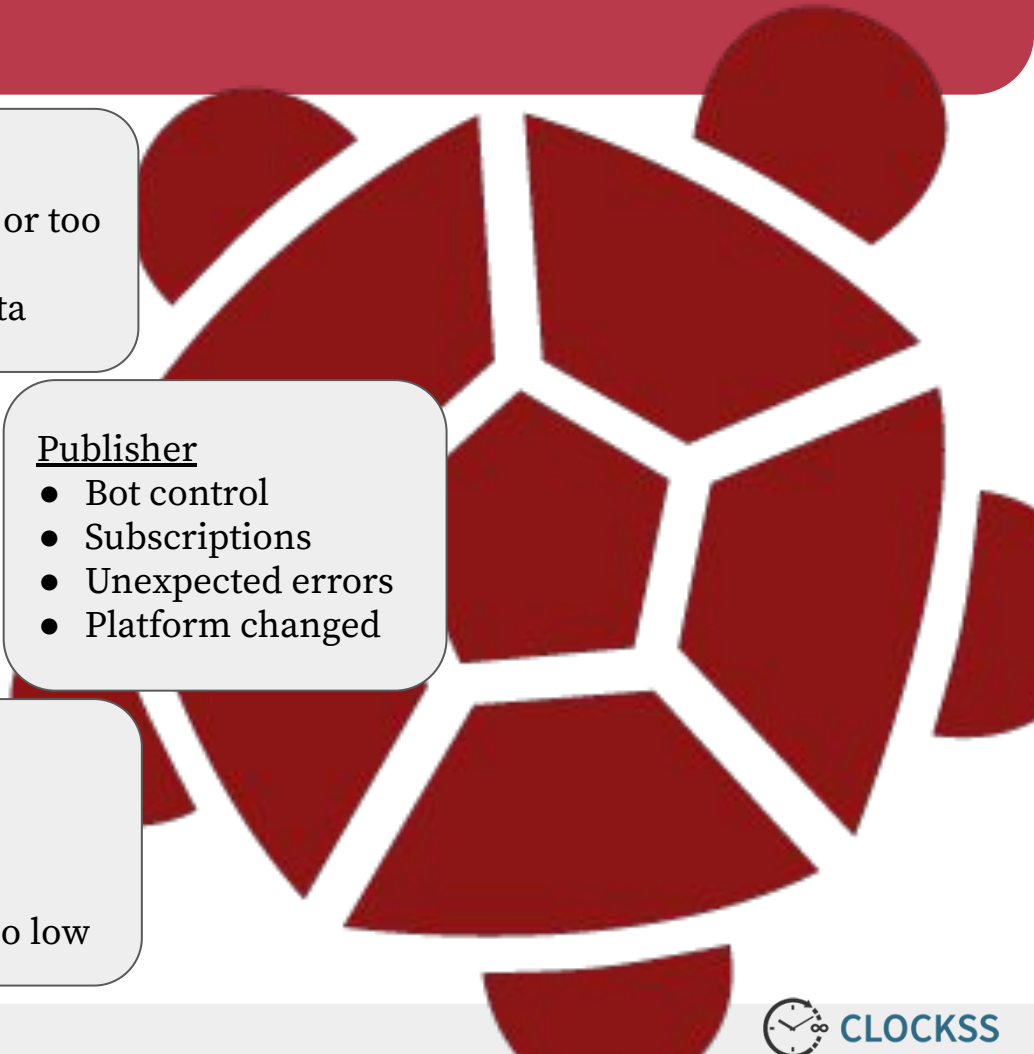
- Incorrect journal parameters
  - Correct parameters.
- Malformed absolute urls (journal.org/www.random.com)
  - Request correction from publisher.
- Malformed link (recursive urls, one-time urls, white space)
  - Request plugin fix.
- More complex issues
  - Ask for help from engineering staff.

Debugging: Hashing < 100%

- Page changes on each browser reload.
  - Variation often transient; wait and retest.
- Page has dynamic content that changes over time or location (ads, related content, in the news, most downloaded, Welcome!)
  - Add a hash filter.
- PDFs are dynamically generated and have 'printed at Stanford'.
  - Add a pdf filter.



Bugs and Resolutions



Content Testing
- Delete extra, or malformed AUs
- Restructure access data

Metadata QA
- Counts too high or too low
- Missing metadata

Bug

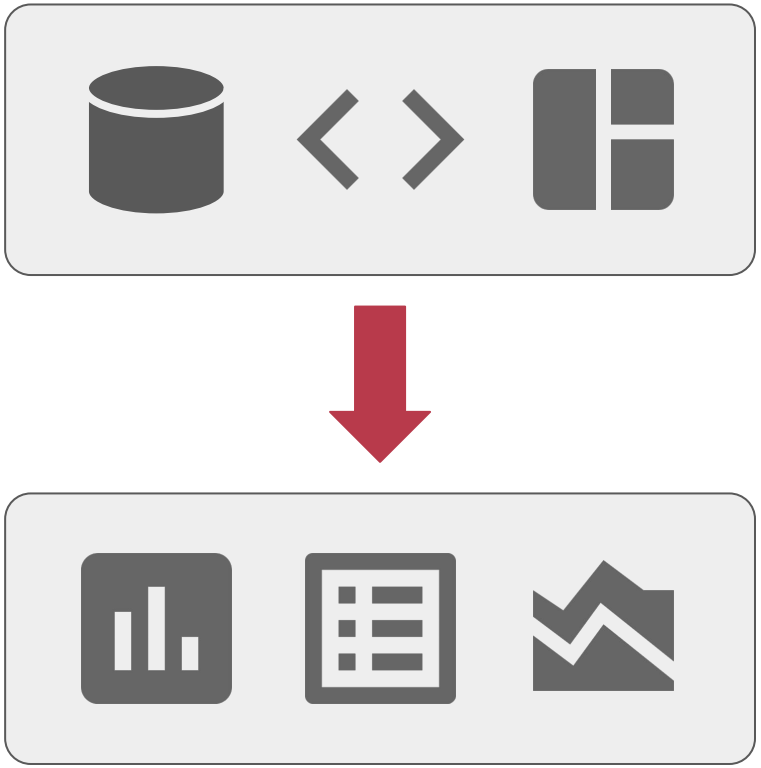
Publisher
- Bot control
- Subscriptions
- Unexpected errors
- Platform changed

Plugin Writers
- Overcrawls or other publisher structure issues

CLOCKSS
- No response from publisher
- List of titles
- Stats too high or too low

Statistics and Reporting

Statistics & Reporting



Data Sources

- Title database files
- Database dumps from nodes
- Reports from JIRA

Reports

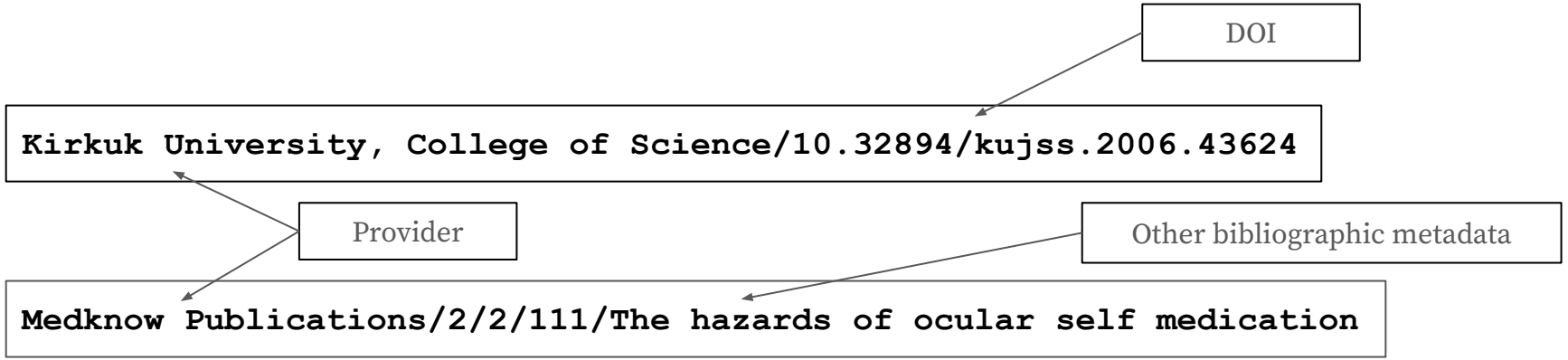
- Weekly Keepers & KBART
- Monthly Reports to CLOCKSS
- Reports uploaded to CLOCKSSface customer portal

Statistics & Reporting

Because we are looking at multiple nodes in our database, it's important to be able to dedupe articles across network nodes.

Our system:

- generates an article key based on available metadata, DOI preferred
- indicates the number of copies of each article by the key



Conclusion

Communication and Documentation Tools

Communication

Zoom
Remote meetings are common with both our team and external

Slack
Internal team communication

Email
Office 365 through Stanford; largely used for external communication

Issue Tracking

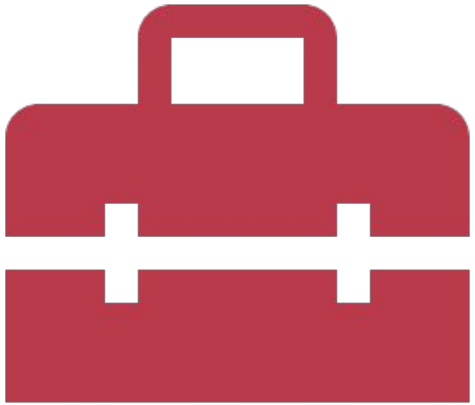
Jira
Internal ticket tracking and work planning

RT
Customer facing ticketing system

Version Control

GitLab and GitHub

Code repositories



Documentation

Internal wiki (Confluence)
Document internal practices; serves as a starting point for new employees

Google Docs
Documentation which can be shared with partners

For more information

For more on CLOCKSS: https://clockss.org/

For more on LOCKSS: https://www.lockss.org/

And feel free to email us with follow up questions! info@lockss.org

Meghan Frazer
Operations Manager
frazerm@stanford.edu

Carmen Cox
Plugin Developer
crc10@stanford.edu

Mary-Ellen Petrich
Digital Preservation Analyst
mpetrich@stanford.edu

Thank you!