On Tool Building and Evaluation of the Archived Web

Mat Kelly
Old Dominion University
Web Science & Digital Libraries Research Group
Department of Computer Science
Norfolk, Virginia USA
mkelly@cs.odu.edu
Who I Am

- PhD Candidate (ADB) of Computer Science
- Defending Dissertation in 2019
- Floridian moving progressively Northward
My Research Topic

- Personal, Private, and Public Web archiving
- Technical perspective involving standards
- Tool Builder to support research
  - Often with a solution seeking a problem
The Origin Topic

- Started off researching wireless sensor networks
- Focus: distributed emergency detection
- Exploratory NesC trilateration implementation

https://github.com/machawk1/alert-codebase
Shift Topics & Labs with My PhD Advisor

(Dr. Michele C. Weigle)
Web Archives? Like Internet Archive?

- Saving pages on the Web of today
  - For exploration and research later
- The Archived Web: A Culturally significant resource
- The Internet Archive (IA) started saving the Web in 1996

But there are **other institutional, public archiving efforts** beyond IA
Digital History on the Web

Now

February 2003

PSU Seminar: On Tool Building and Evaluation of the Archived Web

February 13, 2019

Mat Kelly
@machawk1
PSU.edu of the past

What did PSU.edu look like in the past?

- 2019
- 2012
- 2010
- 2006
- 2001
Multiple archival efforts (3 of many)
First brush with Web Archiving beyond IA

- Tasked to revamp and maintain ArchiveFacebook
  - Mozilla Firefox extension/add-on
- Provided mechanism to allow preservation of user’s Facebook contents
- Created browser-accessible cache of FB web pages
Data Liberation vs. WYSIWYG

Facebook Native Profile Download

Mat Kelly

- **Sex:** Male
- **Birthday:** 1/18/1982
- **Relationship Status:** Married - Melissa Kelly
- **Family:**
  - Jennifer Kelly Price (sister)
  - Melissa Kelly (sister)
  - Michelle Glaser Kelly (mother)
  - Jill Cramer (cousin)
  - Elle Cramer (cousin)
  - Cramer to Jesus (cousin)
  - Steve Glaser (cousin)
  - Kevin Glaser (uncle)
  - Sharon Robbins Glaser (aunt)
  - Carol Bartal (aunt)
  - Ellen Kelly (grandmother)
  - Joyce Baker (aunt)
  - Kelly Baker (cousin)
  - Brian Kelly (uncle)
  - Renee Kelly Szczepanek (cousin)
  - Rebecca Stage (cousin)

- **Email:** me@matkelly.com
- **Facebook Profile:** http://www.facebook.com/profile.php?

Archive Facebook Archiving Session Result
Research Software Beyond my Use Case

- Rapid prototyping
- Public releases of software
- Open source, permissively licensed (GPL or MIT)

Rationales for Tool Build:
- Data generation for further experimentation
- Medium melding and merging (e.g., live & archived Web)
- Exploration on the dynamics of previously unpreserved
Lessons Learned

- Site-specific scrapers are fragile
- Little guidance on the Web on Archival Tool Building
- Testing was ad hoc and laborious (moving target) but effective
- Created Framework for MS Thesis
  - Made these sort of tools more robust and adaptive
Site-Agnostic Preservation

- Preserve everything you see!
- Created files that adhere to standard ISO28500 (Web ARCHive) format
- Enable individuals to preserve any Web page from their browser

github.com/machawk1/warcreate


PSU Seminar: On Tool Building and Evaluation of the Archived Web

February 13, 2019
WARCreate for Google Chrome

- Facilitate preservation through familiar viewport (the browser)
- Extension for Google Chrome
  - Predated WebExtensions standard API
- Easy usage:
  - One-click, current webpage → WARC
- Acts as a “buffer” until commanded to create WARC

github.com/machawk1/warcreate
Archiving the Previously Unarchivable

- Target audience are for users that won’t go to CLI
- Leveraging browser medium was novel and facilitated consistency
Initial limitations

- Interacting with the File System was Limited
  - This was pre-HTML 5 File API
- Initial idea was to have Server-Side replay to also mitigate file limitations
  - This spun off “Web Archiving Integration Layer”
- As File APIs evolved, a “Server” was no longer needed for WARC generation
High Level Overview of WARC format

Concatenated records consisting of:

- WARC Headers
- WARC Payload
  - HTTP Headers
  - HTTP Payload

Information on HTTP request, response, DNS, metadata, general resources
Web Archiving Integration Layer (WAIL)

- Written in Python, compiled to native application
  - initially OS X, Windows, and Linux
- Bundled and preconfigured “Institutional Grade” archiving tools
  - Heritrix (archival grade Web crawler)
  - OpenWayback (Web archive replay system)
- Again, simple interfaces to facilitate usage
Software that uses WARCsf

Writers/Crawlers

HERITRIX
WARCreate
GNU Wget

Readers/Replay Engines

OpenWayback
pywb
ipwb
...
IA’s Web Archives, stored in WARCds, use same tools

WayBack Machine

OpenWayback

HERITRIX
Studying the Archived Web
Beyond Tools and Formats
Study of Archiving Difficulties

- An initial examination of large Web archives
  - cf.live Web
- Which things are hard to preserve?


PSU Seminar: On Tool Building and Evaluation of the Archived Web

February 13, 2019
Personalized Pages in the Archives

- An initial examination of large Web archives cf. live Web
- Some preserved things are personalized

Existing Tools’ Capabilities

Punchline:

● Preservation tools lag in capability cf. Web browsers
● How well do archiving tools perform?
Representations can be **Linked** in time

HTT{P Link
original

HTTP GET
Accept-Datetime: 2012

HTTP GET
Accept-Datetime: 2006

HTTP Link
timegate

HTTP Link
original

HTTP Link
original

HTTP Link
original
Memento provides relations


Memento Request Example

HTTP Request
- Accept-Datetime: Wed, 02 Aug 2017 23:15:00 GMT
- GET: http://web.archive.org/web/http://www.psu.edu

Request psu.edu at Sept 11, 2001 at 9am EST
Memento Request Example

HTTP Request
- **Accept-Datetime:** Wed, 02 Aug 2017 23:15:00 GMT
- **GET:** http://web.archive.org/web/http://www.psu.edu

Request `psu.edu` at Sept 11, 2001 at 9am EST

HTTP Response (302)
- **Memento-Datetime:** Wed, 02 Aug 2017 23:18:04 GMT
- **Location:** http://web.archive.org/web/20170802231804/http://www.psu.edu
- **Link:**
Background: Dereferencing a TimeMap at URI-T

- Date-based pagination
- Other formats for TimeMap

Request URI-T

TimeMap

URI-R

URI-G

URI-M

URI-T
Memento “Damage” Metric

- Not all missing resource are created equal
- Multiple studies establishing metric
- Evaluated through Mechanical Turk

---

**Abstract**

Web archives are important for preserving the digital history of the internet. This entails annotating pages that may become obsolete, broken or missing. These resources are often hierarchically structured, and the process of detecting and cataloguing them can be time-consuming. Therefore, it is necessary to estimate the quality of a web archive by measuring the severity of the damage caused by missing resources.

We propose a new metric called the Memento Damage Metric. This metric is based on the idea that some missing resources are more impactful than others. We measured the damage caused by missing resources based on the number of broken links and the number of missing resources.

**Categories and Subject Descriptors**

- H.3.3 [Information Storage and Retrieval]: Digital Libraries
- H.3.4 [Information Storage and Retrieval]: Information Search and Retrieval
- I.3.7 [Computer Graphics]: Image Representation - Measurement

**Keywords**

- Web architecture
- APTT
- Web archiving
- Edged Processing
- Metrics

---

**Introduction**

Not all Mementos are created equal: Measuring the Impact of Missing Resources

Justin F. Brindza, Justin F. Brindza, Justin F. Brindza, Justin F. Brindza

ABSTRACT

Web archives are important for preserving the digital history of the internet. This entails annotating pages that may become obsolete, broken or missing. These resources are often hierarchically structured, and the process of detecting and cataloguing them can be time-consuming. Therefore, it is necessary to estimate the quality of a web archive by measuring the severity of the damage caused by missing resources.

We propose a new metric called the Memento Damage Metric. This metric is based on the idea that some missing resources are more impactful than others. We measured the damage caused by missing resources based on the number of broken links and the number of missing resources.
Impact of Missing Resources
Damage in the eyes of Mechanical Turkers

- $m_0$: live Web
- $m_1$: comic removed
- $m_2$: two logo images removed

The turkers selected $m_0$ as the preferred memento 81% of the time, and more consistently for larger $\Delta M_m$ values.

<table>
<thead>
<tr>
<th>$\Delta M_m$</th>
<th>5-0</th>
<th>4-1</th>
<th>3-2</th>
<th>2-3</th>
<th>1-4</th>
<th>0-5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>0.8</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>0.3</td>
<td>15</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
</tr>
<tr>
<td>0.2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>0.1</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td></td>
<td>1</td>
<td>0.29</td>
</tr>
<tr>
<td>0.0</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td>0.22</td>
</tr>
<tr>
<td>Total</td>
<td>0.58</td>
<td>0.23</td>
<td>0.09</td>
<td>0.09</td>
<td>0.00</td>
<td>0.02</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Impact of JavaScript on Archivability

- Missing JavaScript has big ramifications
- Content Complexity (CC) measure
- URIs shared over Twitter and from Archive-It collection
- Evaluated WebCitation, wget, and the Heritrix
- 4.2% of the Twitter collection is perfectly archived by all of these tools
- 34.2% of the Archive-It collection is perfectly archived.
Missing resources a direct result of JS usage

www.doc.alabama.gov
perfectly archived in time

cmt.com
Impact of JS evident in missing resources
Problem was getting worse

Mobile Mink

- Android Application
  - Few web archiving offerings in this realm
  - Leveraged Native "Share" feature for archival lookup
- Identified and associated archived mobile representations

**ABSTRACT**

Web surfing mobile app `Mobile Mink` enriches `Android` mobile phone data on archival web. `Mobile Mink` provides mobile and desktop UMA and desktop presentation. `Mobile Mink` also allows sharing a portion of `Web Archive` (WAB) for quick lookup.

Categories and Subject Descriptions

H.31 (Information Storage and Retrieval) Digital Libraries

General Terms

Design, Experimentation, Measurement

Keywords

Web Archiving, Digital Preservation, Memento, Timelines

1. INTRODUCTION

`Mobile Mink` is a mobile framework for Google Chrome that leverages the Memento technology to persist the mobile user experience and the content that the user engages with. The mobile app facilitates the retrieval of web content that the user has previously visited on their mobile device.

2. AGGREGATE TIMELAPS

`Mobile Mink` is an Android application that is currently undergoing development and has been released for download via the Google Play store. The app allows users to share web content in their timeline, providing a way to view past websites and web pages they have visited. The app also includes features for managing and organizing the user’s history, allowing users to easily access and manage their web archiving data.

3. CONCLUSION

`Mobile Mink` is a promising tool for mobile archiving, offering a unique approach to preserving mobile web content. Its integration with Google Chrome provides a seamless experience for users, allowing them to quickly and easily access archived web pages from their devices. The app’s focus on sharing and organizing web content makes it a valuable tool for both personal and professional use, offering a new way to manage and access web content.

4. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under award number...
WAIL reimagined

- Archive from the desktop with higher fidelity than conventional archiving tools

WAIL: Collection-Based Personal Web Archiving

John A. Berlin, Mat Kelly, Michael I. Nahas, Michael C. Wiader
Old Dominion University, Department of Computer Science, Norfolk, VA, 23529 0

Copyright © 2019, Old Dominion University. All rights reserved.

Abstract

Web Archiving Integration Layer (WAIL) is a desktop application written in Python for integrating electron and ElectronJS into web archiving tools. WAIL enables a reimagined personal web archiving tool that allows users to easily and efficiently manage a personal web archive. WAIL does this by integrating native web technologies (e.g., Jupyter, Chromium) to create a new, powerful web browsing experience that can be used to interact with the archived web. WAIL also provides a simple, user-friendly interface for managing and extending the capabilities of existing web browsers, allowing users to easily access and interact with their personal web archive.

Keywords

Personal Web Archiving

1 Introduction

Web Archiving: Integrating Layer (WAIL) is a desktop application written in Python for integrating electron and ElectronJS into web archiving tools. WAIL enables a reimagined personal web archiving tool that allows users to easily and efficiently manage a personal web archive. WAIL does this by integrating native web technologies (e.g., Jupyter, Chromium) to create a new, powerful web browsing experience that can be used to interact with the archived web. WAIL also provides a simple, user-friendly interface for managing and extending the capabilities of existing web browsers, allowing users to easily access and interact with their personal web archive.

Introduction

Web Archiving: Integrating Layer (WAIL) is a desktop application written in Python for integrating electron and ElectronJS into web archiving tools. WAIL enables a reimagined personal web archiving tool that allows users to easily and efficiently manage a personal web archive. WAIL does this by integrating native web technologies (e.g., Jupyter, Chromium) to create a new, powerful web browsing experience that can be used to interact with the archived web. WAIL also provides a simple, user-friendly interface for managing and extending the capabilities of existing web browsers, allowing users to easily access and interact with their personal web archive.

2 WAIL

WAIL is a collection-based personal web archiving tool that integrates native web technologies (e.g., Jupyter, Chromium) into a user-friendly, web-based interface. WAIL provides users with a powerful tool for managing and accessing their personal web archive. WAIL allows users to easily access and interact with their personal web archive, providing a powerful tool for managing and accessing their personal web archive.
<table>
<thead>
<tr>
<th>Code</th>
<th>Python</th>
<th>HTML, JavaScript, Electron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archival organization</td>
<td>Single Collection</td>
<td>Multiple Collections</td>
</tr>
<tr>
<td>User interface</td>
<td>System Native</td>
<td>Material</td>
</tr>
<tr>
<td>Archival Crawler</td>
<td>Heritrix</td>
<td>Heritrix, node-warc</td>
</tr>
<tr>
<td>Archival Replay</td>
<td>OpenWayback</td>
<td>pywb</td>
</tr>
<tr>
<td>Release</td>
<td>macOS, Windows</td>
<td>macOS, Windows, Linux</td>
</tr>
<tr>
<td>Source</td>
<td><a href="github.com/machawk1/wail">github.com/machawk1/wail</a></td>
<td><a href="github.com/n0tan3rd/wail">github.com/n0tan3rd/wail</a></td>
</tr>
</tbody>
</table>

PSU Seminar: On Tool Building and Evaluation of the Archived Web

February 13, 2019

Mat Kelly
@machawk1
Conicalization’s effects over time

URI coalescence considered harmful for archives
Many URI-Ms are actually redirects

<table>
<thead>
<tr>
<th>host</th>
<th>% 3XX</th>
<th>% 200</th>
<th>$M_{TM}$</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>google</td>
<td>84.89</td>
<td>15.11</td>
<td>695,525</td>
<td>0.178</td>
</tr>
<tr>
<td>yahoo</td>
<td>88.16</td>
<td>11.83</td>
<td>418,896</td>
<td>0.134</td>
</tr>
<tr>
<td>sourceforge</td>
<td>73.34</td>
<td>26.63</td>
<td>31,408</td>
<td>0.363</td>
</tr>
<tr>
<td>instagram</td>
<td>67.32</td>
<td>32.65</td>
<td>55,228</td>
<td>0.485</td>
</tr>
<tr>
<td>vimeo</td>
<td>57.04</td>
<td>42.94</td>
<td>199,262</td>
<td>0.752</td>
</tr>
<tr>
<td>cnn</td>
<td>49.97</td>
<td>50.01</td>
<td>87,148</td>
<td>1.001</td>
</tr>
<tr>
<td>wikipedia</td>
<td>44.62</td>
<td>55.19</td>
<td>25,973</td>
<td>1.240</td>
</tr>
<tr>
<td>whitehouse</td>
<td>44.57</td>
<td>55.24</td>
<td>26,006</td>
<td>1.243</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>host</th>
<th>% 3XX</th>
<th>% 200</th>
<th>$M_{TM}$</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>stanford</td>
<td>62.14</td>
<td>37.84</td>
<td>19,309</td>
<td>0.609</td>
</tr>
<tr>
<td>princeton</td>
<td>60.10</td>
<td>39.88</td>
<td>9,355</td>
<td>0.663</td>
</tr>
<tr>
<td>columbia</td>
<td>48.01</td>
<td>51.88</td>
<td>9,882</td>
<td>1.082</td>
</tr>
<tr>
<td>harvard</td>
<td>33.91</td>
<td>65.96</td>
<td>7,699</td>
<td>1.948</td>
</tr>
<tr>
<td>caltech</td>
<td>33.13</td>
<td>66.86</td>
<td>5,474</td>
<td>2.017</td>
</tr>
<tr>
<td>mit</td>
<td>26.57</td>
<td>73.24</td>
<td>6,379</td>
<td>2.763</td>
</tr>
<tr>
<td>gatech</td>
<td>26.03</td>
<td>73.94</td>
<td>3,907</td>
<td>2.841</td>
</tr>
<tr>
<td>ufl</td>
<td>24.76</td>
<td>75.23</td>
<td>4,927</td>
<td>3.038</td>
</tr>
<tr>
<td>vt</td>
<td>23.07</td>
<td>76.92</td>
<td>4,061</td>
<td>3.334</td>
</tr>
<tr>
<td>lsu</td>
<td>15.06</td>
<td>84.93</td>
<td>2,974</td>
<td>5.638</td>
</tr>
<tr>
<td>nsu</td>
<td>13.82</td>
<td>86.00</td>
<td>1,208</td>
<td>6.233</td>
</tr>
<tr>
<td>odu</td>
<td>9.727</td>
<td>90.27</td>
<td>1,727</td>
<td>9.279</td>
</tr>
<tr>
<td>tcc</td>
<td>5.429</td>
<td>94.57</td>
<td>884</td>
<td>17.41</td>
</tr>
</tbody>
</table>

InterPlanetary Wayback (ipwb)

- Personal archives are more resilient when propagated.
- How much does it cost to have resilient personal archives?

PSU Seminar: On Tool Building and Evaluation of the Archived Web
February 13, 2019

Mat Kelly
@machawk1
- Persistence of archived web data dependent on resilience of organization and availability of data
- Remove massive redundancy in web archive files of exact duplicate content
- Determine feasibility of pushing WARCs into IPFS
ipwb Base Dynamics

- IPWB CDXJs may be transferred for our users’ replay
  
  ![Diagram showing WARC, ipwb, CDXJ, and user](image)

- CDXJ-by-hash recursive fetch/replay
  - Share hash of CDXJ then `$ ipwb replay hash` to replicate experience

  ![Diagram showing Push CDXJ to IPFS](image)

---

PSU Seminar: On Tool Building and Evaluation of the Archived Web

February 13, 2019

Mat Kelly
@machawk1
WARC Store

1. Indexer
   extract HTTP headers+payloads

CDXJ

IPFS Store

Replay
combine header+payload

@machawk1
PSU Seminar: On Tool Building and Evaluation of the Archived Web

February 13, 2019

@machawk1
PSU Seminar: On Tool Building and Evaluation of the Archived Web

February 13, 2019

@machawk1

Indexer
extract HTTP
headers+payloads

CDXJ

index record(s)

URI, [datetime]

1

2

3

Replay
combine
header+payload

WARC Store

IPFS Store
PSU Seminar: On Tool Building and Evaluation of the Archived Web
February 13, 2019

Mat Kelly
@machawk1

WARC Record extraction to CDXJ

ipwb.example.com)/20160905022013 {"locator":
"urn:ipfs/QmcN9eWwRF73dZj5BgT4x8jeEcFrXurX1hot8QwCbM19PB/Qmczh9YnB41ptPeqxcaTZA4aMmuNUswTLTWzXtvnbp9sL",
"mime_type": "text/html", "status_code": "200"
}

ipwb.example.com/style.css 20160905022013 {"locator":
"urn:ipfs/QmU1k71bT6ibZBSdxBL35cQXwovTih8cTB4CXfrjyMfZX/E/QmbvUAo9U31wSdvARjvbPeVBTAWCjN1kyPhQ4ho3n8TAZo",
"mime_type": "text/css", "status_code": "200"
}

ipwb.example.com)/ipwb.png 20160905022013 {"locator":
"urn:ipfs/QmTjfMxFGvbP4nwF0q3tNYDPW6gC99i5nrq5XSw6QrvHa/QmYMKZbnk53ukuPJJirahJHGevCCy2afLyePRdX38TukFUwd",
"mime_type": "image/png", "status_code": "200"
}

ipwb.example.com)/fileduration.png 20160905022013 {"locator":
"urn:ipfs/QmaCj6LNngxwxqxlMmp1xCyxcwDlt2Uzqf8gCG6bVyQppYCY/QmdgtMcGprTF8bqv7ytgMwt0i5BhRxfuvBjD6Vj2U7ohz1",
"mime_type": "image/png", "status_code": "200"
}

ipwb.example.com)/filesize.png 20160905022013 {"locator":
"urn:ipfs/QmNPjrSVY31oGDooMA18ZDHNfKlnEg3j5gRj1dFdrqmS4/Qmb4heB8PU58nkWt6w5tBgMfpeLTKuU7iuxg9tFdoPsF1B",
"mime_type": "image/png", "status_code": "200"}
IPFS multihashes in IPWB CDXJs

com,example)/index.html 20170301192639 {"locator": "urn:ipfs/QmPdyY6Pm66iWtGpTc7PqK11hvsnYSKMVL57G69RiNjGcm/QmNZ6mKSAXAmXEocQj5gT4y4kdcr5D2C173ubWJ6PSKEZ", "mime_type": "text/html", "status_code": "200"}

com,example)/images/frog.png 20170301192639 {"locator": "urn:ipfs/QmUeko8zM7Xanwz6F9GtRH4rLAi4Poj3EMECGsci3BRQfs/QmPhMnX74cwqx2xgj9d3N3gTra8CzafXwSbUwU8xagMfqR", "mime_type": "image/png", "status_code": "200"}
Content Addressing

http://foo.com/spaceDog.jpg

http://example.org/yuri.jpg

$ ipfs cat QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4 > doge.jpg
Methodology - IPWB WARC indexing

- `warc-response` record body extracted into temp files
  - HTTP header and entity body (payload) separated
  - Response metadata (e.g., datetime) retained
- Temp files pushed into IPFS via locally running daemon
  - Two IPFS hashes (for header and payload) returned
- CDXJ record created representing `warc-response` contents
  - Contains URI-R, archived HTTP status, encoded IPFS hashes
Methodology - Replaying Archives

- Extension of pywb API to read CDXJ files
- On encountering IPFS URN, fetch \texttt{warc-response temp} files from IPFS using local daemon
  - This may occur on a separate machine using a separate daemon
- With WARC contents fetched, replay contents using pywb where the locator value in the CDXJ is used to dereference the temp files pulled from IPFS
CDXJ in ipwb

```
SURT_URI DATETIME {
    "id": "WARC-Record-ID",
    "url": "ORIGINAL_URI",
    "status": "3-DIGIT_HTTP_STATUS",
    "mime": "Content-Type",
    "locator": "urn:ipfs/HEADER_DIGEST/PAYLOAD_DIGEST"
}
```
A Framework for Aggregating Private and Public Web Archives

- Mitigate outstanding issues in Web archiving beyond public scope.
- High-level of dissertation topic
- Introduced the “Mementy Framework”
Desire: Include Other Non-Aggregated Archives
Archiving More Archives Provides a Better Picture of the Web
Query Precedence & Short-Circuiting

- May give priority to archive relevancy.
- Series halt when threshold met.

“Check private archives first. **Iff** you find no captures, only *then* check the public archives.
Aggregation with **Access Regulation**

HTTP 401

Consult URI-P

memento

META

AGGREGATE

ALICE’S private captures

✓

Bank of America

Consult

Locked

-figure with access regulation and aggregation concepts-

---

PSU Seminar: On Tool Building and Evaluation of the Archived Web

February 13, 2019
OAuth2-based tokenization patterns
...with offloading of the procedure from the archives
Evaluation Through Implementation

- Extend for client-side archival specification
- Exhibit features of an MMA
- Regulate access to Private Web archives
- Facilitate archival negotiation in more dimensions
On Tool Building and Evaluation of the Archived Web

Open Source Web Archiving Tools

More details of studies:

- Measuring the Impact of Missing Resources ([Conf](#), [Journal](#))
- Impact of URI Canonicalization ([Conf](#), [arXiv](#))
- Archivability Over Time ([Conf](#), [arXiv](#))
- Impact of JS on Archivability ([Journal](#))
- Personalization in Web Archives ([article](#))

Slides available at [http://matkelly.com/psu](http://matkelly.com/psu)