To the Rescue of the Orphans of Scholarly Communication

Herbert Van de Sompel
Los Alamos National Laboratory  @hvdsomp
http://orcid.org/0000-0002-0715-6126

Michael L. Nelson
Old Dominion University  @phonedude_mln
http://orcid.org/0000-0003-3749-8116

Martin Klein
Los Alamos National Laboratory  @mart1nkle1n
http://orcid.org/0000-0003-0130-2097

The project is funded by the Andrew W. Mellon Foundation
Outline

• Problem statement

  *Scholarly artifacts are everywhere on the web, and are not systematically archived*

• Project perspective

  *Capturing artifacts using an institutional & web archiving paradigm*

• Object capture flow:
  • Step 1: Discovering a researcher’s web identities
  • Step 2: Discovering artifacts per web identity
  • Step 3: Determining the web boundary per artifact
  • Step 4: Capturing resources in the artifacts’ web boundary
Scholarship is Evolving

• The research process, not just its outcome, is becoming visible … on the web

• Massive extension of the scholarly record with an enormous variety of novel artifacts

• The artifacts are heterogeneous, dynamic, compound, inter-related and distributed across the web

• The artifacts are often hosted on common web platforms that are not dedicated to scholarship
Bianca Kramer & Jeroen Bosman. 101 Innovations in Scholarly Communication
https://innoscholcomm.silk.co/
The Evolving Scholarly Record

Brian Lavoie et al. (2014) The Evolving Scholarly Record
Web Platforms Record Scholarship

• Increasingly, common web platforms are used for scholarship
  • GitHub, Wikis, Wordpress, etc.

• Many of these platforms have desirable characteristics
  • Versioning
  • Time stamping
  • Social embedding

• But, these platforms record artifacts rather than archive them

Herbert Van de Sompel & Andrew Treloar (2014) A Perspective on Archiving the Scholarly Web
Recording is not Archiving

“GitHub reserves the right at any time and from time to time to modify or discontinue, temporarily or permanently, the Service (or any part thereof) with or without notice.”

Recording is not Archiving

Google Open Source Blog
News about Google’s open source student programs and software releases

Bidding farewell to Google Code
Thursday, March 12, 2015

When we started the Google Code project hosting service in 2006, the world of project hosting was limited. We were worried about reliability and stagnation, so we took action by giving the open source community another option to choose from. Since then, we’ve seen a wide variety of better project hosting services such as GitHub and Bitbucket bloom. Many projects moved away from Google Code to those other systems. To meet developers where they are, we ourselves migrated nearly a thousand of our own open source projects from Google Code to GitHub.

https://opensource.googleblog.com/2015/03/farewell-to-google-code.html
## Recording versus Archiving

<table>
<thead>
<tr>
<th>Recording</th>
<th>Archiving</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-term</td>
<td>Longer-term</td>
</tr>
<tr>
<td>No guarantees provided</td>
<td>Attempt to provide guarantees</td>
</tr>
<tr>
<td>Write many/read many</td>
<td>Write once/Read many</td>
</tr>
<tr>
<td>Scholarly process</td>
<td>Scholarly record</td>
</tr>
</tbody>
</table>

Herbert Van de Sompel & Andrew Treloar (2014) A Perspective on Archiving the Scholarly Web
Meet Some New School Researchers

Ian Milligan
University of Waterloo

Mark Matienzo
Stanford University
Meet Some New School Researchers

https://ianmilligan.ca/
https://www.slideshare.net/IanMilligan1
https://github.com/ianmilligan1

Ian Milligan
Meet Some New School Researchers

Mark Matienzo

http://matienzo.org/
https://www.slideshare.net/anarchivist/
https://github.com/anarchivist
https://osf.io/tgr4k/
https://www.drupal.org/user/380762
Ian’s SlideShare Artifact: 0 Mementos

https://www.slideshare.net/IanMilligan1/resaw-geo-cities
http://timetravel.mementoweb.org/list/20140513211653/https://www.slideshare.net/IanMilligan1/resaw-geo-cities
Ian’s GitHub Artifact: 1 Memento

https://github.com/ianmilligan1/Historian-WARC-1
The Scholarly Orphans Project

- Funded by the Andrew W, Mellon Foundation
  - Los Alamos National Laboratory & New Mexico Consortium
  - Old Dominion University
  - 04/2016 - 03/2019

- How to capture scholarly orphans for long-term archiving?

- Project explores a paradigm inspired by web archiving
  - Scale of the problem
  - Bilateral agreements with most web portals unlikely

- Project explores an institution driven paradigm
  - Institution should be interested in capturing the artifacts its scholars deposit on the web
An Institutional & Web Archiving Perspective
Related Work

• LOCKSS
  • Web crawling approach
  • Focused on journal literature

• Archive-It
  • On-demand, subscription-based web archiving
  • Not focused on scholarly orphans

• Institutional repository
  • Capture an institution’s output
  • Focused on manual upload (of journal literature)

• The Locker Project
  • Capture an individual’s web presence
  • Not focused on scholarly orphans
Capture Flow

Step 1: Discovery of a Researcher’s Web Identities
- Algorithmic Approach
- Web Identity Registry

Web Identities

Step 2: Discovery of Artifacts per Web Identity
- Algorithmic Approach
- Notifications
- Artifact Registry

URIs

Step 3: Determination of Web Boundary per Artifact
- Algorithmic Approach
- Signposting

URIs

Step 4: Capture Resources in Artifact Boundary
- Capture Tools
- Capture Quality Evaluation
- Capture Authenticity Verification

Archive

@hvdsomp, @phonedude_mln, @mart1nkle1n
CNI Spring 2017, Albuquerque, NM, 3 Apr 2017
Capture Flow – Step 1

Step 1: Discovery of a Researcher’s Web Identities
- Algorithmic Approach
- Web Identity Registry

Web Identities

Step 2: Discovery of Artifacts per Web Identity
- Algorithmic Approach
- Notifications
- Artifact Registry

Step 3: Determination of Web Boundary per Artifact
- Algorithmic Approach
- Signposting

Step 4: Capture Resources in Artifact Boundary
- Capture Tools
- Capture Quality Evaluation
- Capture Authenticity Verification

URIs

Archive

@hvdsomp, @phonedude_mln, @mart1nkle1n
CNI Spring 2017, Albuquerque, NM, 3 Apr 2017
Algorithmic Discovery of Web Identities

Discovery of Web Identities via a Registry: ORCID

Ian Milligan
http://orcid.org/0000-0002-1470-7723

Mark Matienzo
http://orcid.org/0000-0003-3270-1306

Martin Klein and Herbert Van de Sompel (2017) Discovering scholarly orphans using ORCID
Ian Milligan’s ORCID

- Web Identities: 0

http://orcid.org/0000-0002-1470-7723
Mark Matienzo’s ORCID

- Web Identities: 3 (homepage, ScopusID, ResearcherID)

http://orcid.org/0000-0003-3270-1306
Mark Matienzo’s Home Page

- URI to GitHub repository, Twitter
- Could be included in ORCID profile

http://matienzo.org/
Discovery of Web Identities via a Registry: ORCID

- Evaluation of ORCID for automatic discovery of Web Identities

- How well does ORCID represent the global community of active researchers?
  - Adoption rate
  - Subject coverage
  - Geo-location coverage

- How well does ORCID score when it comes to listing Web Identities?
ORCID - Adoption Rate

- ORCIDs total
- ORCIDs with given names
- ORCIDs with first names
- ORCIDs with works
- ORCIDs with affiliations
- ORCIDs with web identities

2013
2014
2015
2016

0
100,000
200,000
300,000
400,000
500,000
600,000
700,000
800,000

ORCIDs total
ORCIDs with given names
ORCIDs with first names
ORCIDs with works
ORCIDs with affiliations
ORCIDs with web identities
## ORCID - Geo-Location Coverage

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>US</td>
<td>16.9%</td>
<td>US</td>
<td>17.1%</td>
<td>CN</td>
<td>19.1%</td>
</tr>
<tr>
<td>2</td>
<td>ES</td>
<td>8.9%</td>
<td>GB</td>
<td>7.1%</td>
<td>US</td>
<td>16.7%</td>
</tr>
<tr>
<td>3</td>
<td>GB</td>
<td>6.4%</td>
<td>ES</td>
<td>6.9%</td>
<td>JP</td>
<td>8.5%</td>
</tr>
<tr>
<td>4</td>
<td>CN</td>
<td>5.6%</td>
<td>CN</td>
<td>5.6%</td>
<td>RU</td>
<td>5.7%</td>
</tr>
<tr>
<td>5</td>
<td>IN</td>
<td>5.5%</td>
<td>BR</td>
<td>5.6%</td>
<td>DE</td>
<td>4.6%</td>
</tr>
<tr>
<td>6</td>
<td>BR</td>
<td>4.4%</td>
<td>IN</td>
<td>5.2%</td>
<td>KR</td>
<td>4.1%</td>
</tr>
<tr>
<td>7</td>
<td>PT</td>
<td>4.3%</td>
<td>IT</td>
<td>4.1%</td>
<td>FR</td>
<td>3.4%</td>
</tr>
<tr>
<td>8</td>
<td>IT</td>
<td>4.3%</td>
<td>AU</td>
<td>3.1%</td>
<td>GB</td>
<td>3.3%</td>
</tr>
<tr>
<td>9</td>
<td>AU</td>
<td>2.8%</td>
<td>PT</td>
<td>3.0%</td>
<td>IN</td>
<td>2.7%</td>
</tr>
<tr>
<td>10</td>
<td>SE</td>
<td>2.6%</td>
<td>RU</td>
<td>2.9%</td>
<td>CA</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

© Creative Commons Attribution-ShareAlike 4.0 License
## ORCID - Geo-Location Coverage

<table>
<thead>
<tr>
<th>Rank</th>
<th>2015 ORCID</th>
<th></th>
<th>2016 ORCID</th>
<th></th>
<th>2013 UNESCO</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Geo</td>
<td>Freq</td>
<td>Geo</td>
<td>Freq</td>
<td>Geo</td>
<td>Freq</td>
</tr>
<tr>
<td>1</td>
<td>US</td>
<td>16.9%</td>
<td>US</td>
<td>17.1%</td>
<td>CN</td>
<td>19.1%</td>
</tr>
<tr>
<td>2</td>
<td>ES</td>
<td>8.9%</td>
<td>GB</td>
<td>7.1%</td>
<td>US</td>
<td>16.7%</td>
</tr>
<tr>
<td>3</td>
<td>GB</td>
<td>6.4%</td>
<td>ES</td>
<td>6.9%</td>
<td>JP</td>
<td>8.5%</td>
</tr>
<tr>
<td>4</td>
<td>CN</td>
<td>5.6%</td>
<td>CN</td>
<td>5.6%</td>
<td>RU</td>
<td>5.7%</td>
</tr>
<tr>
<td>5</td>
<td>IN</td>
<td>5.5%</td>
<td>BR</td>
<td>5.6%</td>
<td>DE</td>
<td>4.6%</td>
</tr>
<tr>
<td>6</td>
<td>BR</td>
<td>4.4%</td>
<td>IN</td>
<td>5.2%</td>
<td>KR</td>
<td>4.1%</td>
</tr>
<tr>
<td>7</td>
<td>PT</td>
<td>4.3%</td>
<td>IT</td>
<td>4.1%</td>
<td>FR</td>
<td>3.4%</td>
</tr>
<tr>
<td>8</td>
<td>IT</td>
<td>4.3%</td>
<td>AU</td>
<td>3.1%</td>
<td>GB</td>
<td>3.3%</td>
</tr>
<tr>
<td>9</td>
<td>AU</td>
<td>2.8%</td>
<td>PT</td>
<td>3.0%</td>
<td>IN</td>
<td>2.7%</td>
</tr>
<tr>
<td>10</td>
<td>SE</td>
<td>2.6%</td>
<td>RU</td>
<td>2.9%</td>
<td>CA</td>
<td>2.1%</td>
</tr>
<tr>
<td>Rank</td>
<td>2015 ORCID Geo</td>
<td>Freq</td>
<td>2016 ORCID Geo</td>
<td>Freq</td>
<td>2013 UNESCO Geo</td>
<td>Freq</td>
</tr>
<tr>
<td>------</td>
<td>----------------</td>
<td>------</td>
<td>----------------</td>
<td>------</td>
<td>----------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>US</td>
<td>16.9%</td>
<td>US</td>
<td>17.1%</td>
<td>CN</td>
<td>19.1%</td>
</tr>
<tr>
<td>2</td>
<td>ES</td>
<td>8.9%</td>
<td>GB</td>
<td>7.1%</td>
<td>US</td>
<td>16.7%</td>
</tr>
<tr>
<td>3</td>
<td>GB</td>
<td>6.4%</td>
<td>ES</td>
<td>6.9%</td>
<td>JP</td>
<td>8.5%</td>
</tr>
<tr>
<td>4</td>
<td>CN</td>
<td>5.6%</td>
<td>CN</td>
<td>5.6%</td>
<td>RU</td>
<td>5.7%</td>
</tr>
<tr>
<td>5</td>
<td>IN</td>
<td>5.5%</td>
<td>BR</td>
<td>5.6%</td>
<td>DE</td>
<td>4.6%</td>
</tr>
<tr>
<td>6</td>
<td>BR</td>
<td>4.4%</td>
<td>IN</td>
<td>5.2%</td>
<td>KR</td>
<td>4.1%</td>
</tr>
<tr>
<td>7</td>
<td>PT</td>
<td>4.3%</td>
<td>IT</td>
<td>4.1%</td>
<td>FR</td>
<td>3.4%</td>
</tr>
<tr>
<td>8</td>
<td>IT</td>
<td>4.3%</td>
<td>AU</td>
<td>3.1%</td>
<td>GB</td>
<td>3.3%</td>
</tr>
<tr>
<td>9</td>
<td>AU</td>
<td>2.8%</td>
<td>PT</td>
<td>3.0%</td>
<td>IN</td>
<td>2.7%</td>
</tr>
<tr>
<td>10</td>
<td>SE</td>
<td>2.6%</td>
<td>RU</td>
<td>2.9%</td>
<td>CA</td>
<td>2.1%</td>
</tr>
</tbody>
</table>
## ORCID - Web Identities

<table>
<thead>
<tr>
<th>Label</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkedIn</td>
<td>7326 (3.8%)</td>
</tr>
<tr>
<td>Researchgate</td>
<td>5306 (2.7%)</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>4976 (2.6%)</td>
</tr>
<tr>
<td>Personal Website</td>
<td>4513 (2.3%)</td>
</tr>
<tr>
<td>Homepage</td>
<td>2916 (1.5%)</td>
</tr>
<tr>
<td>Academia.edu</td>
<td>2027 (1.0%)</td>
</tr>
<tr>
<td>Research Gate</td>
<td>1750 (0.9%)</td>
</tr>
<tr>
<td>UCL IRIS Profile</td>
<td>1657 (0.9%)</td>
</tr>
<tr>
<td>Home Page</td>
<td>1597 (0.8%)</td>
</tr>
<tr>
<td>Twitter</td>
<td>1487 (0.8%)</td>
</tr>
<tr>
<td>LinkedIn Profile</td>
<td>1433 (0.7%)</td>
</tr>
<tr>
<td>KTH Profile</td>
<td>1379 (0.7%)</td>
</tr>
<tr>
<td>SISIUS</td>
<td>1334 (0.7%)</td>
</tr>
<tr>
<td>ID Dialnet</td>
<td>1304 (0.7%)</td>
</tr>
<tr>
<td>ID Personal SICA</td>
<td>1286 (0.7%)</td>
</tr>
<tr>
<td>Personal</td>
<td>1261 (0.6%)</td>
</tr>
<tr>
<td>Web de la Universidad</td>
<td>1156 (0.6%)</td>
</tr>
<tr>
<td>Personal Webpage</td>
<td>1103 (0.6%)</td>
</tr>
<tr>
<td>Blog</td>
<td>1042 (0.5%)</td>
</tr>
</tbody>
</table>
## ORCID - Web Identities

<table>
<thead>
<tr>
<th>Label</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkedIn</td>
<td>7326 (3.8%)</td>
</tr>
<tr>
<td>Researchgate</td>
<td>5306 (2.7%)</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>4976 (2.6%)</td>
</tr>
<tr>
<td>Personal Website</td>
<td>4513 (2.3%)</td>
</tr>
<tr>
<td>Homepage</td>
<td>2916 (1.5%)</td>
</tr>
<tr>
<td>Academia.edu</td>
<td>2027 (1.0%)</td>
</tr>
<tr>
<td>Research Gate</td>
<td>1750 (0.9%)</td>
</tr>
<tr>
<td>UCL IRIS Profile</td>
<td>1657 (0.9%)</td>
</tr>
<tr>
<td>Home Page</td>
<td>1597 (0.8%)</td>
</tr>
<tr>
<td>Twitter</td>
<td>1487 (0.8%)</td>
</tr>
<tr>
<td>LinkedIn Profile</td>
<td>1433 (0.7%)</td>
</tr>
<tr>
<td>KTH Profile</td>
<td>1379 (0.7%)</td>
</tr>
<tr>
<td>SISIUS</td>
<td>1334 (0.7%)</td>
</tr>
<tr>
<td>ID Dialnet</td>
<td>1304 (0.7%)</td>
</tr>
<tr>
<td>ID Personal SICA</td>
<td>1286 (0.7%)</td>
</tr>
<tr>
<td>Personal</td>
<td>1261 (0.6%)</td>
</tr>
<tr>
<td>Web de la Universidad</td>
<td>1156 (0.6%)</td>
</tr>
<tr>
<td>Personal Webpage</td>
<td>1103 (0.6%)</td>
</tr>
<tr>
<td>Blog</td>
<td>1042 (0.5%)</td>
</tr>
</tbody>
</table>
Discovery of Web Identities via a Registry: ORCID

- Adoption rate is increasing
- Subject coverage is focused, does not cover all disciplines equally
- Geo-Location coverage is good but not quite representative
- Web Identity coverage is poor; not usable for our purpose in its current state
Capture Flow – Step 2

Step 1: Discovery of a Researcher’s Web Identities
- Algorithmic Approach
- Web Identity Registry

Web Identities

Step 2: Discovery of Artifacts per Web Identity
- Algorithmic Approach
- Notifications
- Artifact Registry

URIs

Step 3: Determination of Web Boundary per Artifact
- Algorithmic Approach
- Signposting

URIs

Step 4: Capture Resources in Artifact Boundary
- Capture Tools
- Capture Quality Evaluation
- Capture Authenticity Verification

Archive

@hvdsomp, @phonedude_mln, @mart1nkle1n
CNI Spring 2017, Albuquerque, NM, 3 Apr 2017
Discovery of Artifacts per Web Identity

- Algorithmic approach

- Scrape artifacts from pages

http://matienzo.org/publications/
Discovery of Artifacts per Web Identity

- Notifications
- Subscribe to portal notifications about a researcher’s new artifacts

https://www.slideshare.net/anarchivist/presentations
Discovery of Artifacts per Web Identity

- Artifact Registry
- 5 artifacts of interest (standards document, reports, book reviews)

http://orcid.org/0000-0003-3270-1306
Capture Flow – Step 3

Step 1: Discovery of a Researcher’s Web Identities
- Algorithmic Approach
- Web Identity Registry

Step 2: Discovery of Artifacts per Web Identity
- Algorithmic Approach
- Notifications
- Artifact Registry

Step 3: Determination of Web Boundary per Artifact
- Algorithmic Approach
- Signposting

Step 4: Capture Resources in Artifact Boundary
- Capture Tools
- Capture Quality Evaluation
- Capture Authenticity Verification

Archive
Determination of Web Boundary per Artifact

http://signposting.org
HTTP Links

http://tools.ietf.org/rfc/rfc5988.txt
HTTP Links

URI-1 describes URI-2
HTTP Links

URI-1

describes

URI-2

HTTP GET

HTTP/1.1 200 OK
Date: Thu, 27 Oct 2016 04:43:28 GMT
Content-Type: application/rdf+xml; charset=UTF-8
Content-Length: 1210
Link: <URI-2>; rel="describes"
Signposting - Publication Boundary Pattern

http://signposting.org/publication_boundary/oxford/
Signposting - Bibliographic Metadata Pattern

http://signposting.org/bibliographic_metadata/springer/
Capture Flow – Step 4

**Step 1: Discovery of a Researcher’s Web Identities**
- Algorithmic Approach
- Web Identity Registry

**Step 2: Discovery of Artifacts per Web Identity**
- Algorithmic Approach
- Notifications
- Artifact Registry

**Step 3: Determination of Web Boundary per Artifact**
- Algorithmic Approach
- Signposting

**Step 4: Capture Resources in Artifact Boundary**
- Capture Tools
- Capture Quality Evaluation
- Capture Authenticity Verification

Arrows indicate the flow of information:
- Web Identities → Web Identities
- URIs → URIs
- URIs → Archive
Challenges Regarding Capturing Web Artifacts

- **Legal**
  - robots.txt
  - Licenses

- **Technical**
  - Capture tools
  - Capture quality
  - Capture authenticity
Legal Challenges re Capturing Artifacts – A Wake-Up Call

SlideShare
• robots.txt unclear, some pages disallowed
• License seems to prohibit archiving

GitHub
• robots.txt unclear, some pages disallowed
• License seems to allow archiving

Drupal
• robots.txt allows relevant URIs
• License seems to prohibit archiving

Open Science Framework
• robots.txt does not disallow crawlers
• License does not mention archiving, individual content may have specific license
Capture Tools Challenges: Mark’s SlideShare

Live

Webrecorder.io

Internet Archive

https://www.slideshare.net/anarchivist/to-hell-with-good-intentions-linked-data-and-the-power-to-name
Capture Tools Challenges: Mark’s GitHub

Live

Webrecorder.io

Internet Archive

https://github.com/rightsstatements/rightsstatements.github.io
Capture Tools Challenges: Mark’s OSF

As the Digital Public Library of America (DPLA) has advanced over the last three years, one of the major challenges it has faced is the lack of effective, widespread tools among would-be data providers for stewarding and syndicating digital content. In March 2015, the Institute of Museum and Library Services (IMLS) awarded $2M to DPLA, Stanford and DuraSpace for the “Hydra-in-a-Box” project to help address this need. This is the largest single grant made in IMLS history, and it was made as part of its recently-announced funding priorities, the National Digital Platform. The goals of the 2.5 year project are to polish and package the robust and flexible Hydra repository software into a “feature complete enough” digital repository, and one that can be readily adopted as a turnkey solution or via cloud-based services. While Hydra has demonstrated versatility and strength as a “front door” for those seeking to assemble their own repository-based applications (or “Hydra head”) projects will accelerate development of critical core features AND also bring making the software easy to get up and running. The ultimate goal of the OPF/OSF national platform for managing digital content—one that equips the library community with a best-of-bred repository for local use, and that links together for networked access (including but not only DPLA). This presents an opportunity to look at the project’s goals, timelines, and early progress. Among these are the international survey of hundreds of repositories about desired features, and requirements gathering to feed the user-centered design process. The project touches on how the partners are consciously co-opting the energy of both the communities.

Live

Internet Archive

Webrecorder.io

https://osf.io/h4ru8/wiki/home/
Capture Quality - How well was this page archived?

• Continuing research on *memento damage*, first published at JCDL 2014

• Premise: simply reporting “9/10 embedded images were archived” is insufficient to describe how well the archive / replay system performed

• Use heuristics from Mechanical Turk testing to approximate human conception of damage, e.g.:
  - increase weight of missing images that are large, or centered in the viewport
  - stylesheets can be important! check for “ugly” results

http://dx.doi.org/10.1109/JCDL.2014.6970187 http://dx.doi.org/10.1007/s00799-015-0150-6
Triptych CSS

“regular” web pages have nearly equal distribution of content over each third of a page

if a CSS is missing AND > 75% of the non-background color is in the left 2/3s of the page, then users consider this damaged
A Memento Damage Service, Python Library, and Docker Image

MementoDamage

How well is your webpage archived?

http://odu.edu/compsci
Damage = 0.024

http://clarkjolley.com
Damage = 0.44014

http://alguard.state.al.us
Damage = 0.96800

Check the damage of your page

http://www.somesite.com

Note: The memento damage calculation will work on live webpages, but was designed to evaluate archived webpages or mementos. Discover mementos using Time Travel

Erika Siregar
http://memento-damage.cs.odu.edu/
https://github.com/erikaris/web-memento-damage
Just a Little Bit of Damage…
Moderate Damage…
Significant Damage…
Ian’s GitHub Memento…

... Has Slight Damage

does not appear to violate the “75% / left-2/3s” rule
Capture Authenticity - *Has this page been tampered with?*

- The days of implicitly trusting Brewster & IA are over
  - the people who brought you fake news will eventually bring you *fake archives*
  - “mo’ archives mo’ problems”
- Premise: use multiple, independent archives to record fixity information from dated observations of mementos
- Plans:
  - blockchain
  - provenance (i.e., a memento of memento != 2 independent mementos)

https://climate.nasa.gov/vital-signs/carbon-dioxide/
Push a Web Page into Multiple Archives

Record Fixity in a Manifest File

Shawn Jones (2016) Mementos In the Raw, Take Two
Publish Manifest to the Web

Manifest Server (e.g., captureproject.org)

manifest file

```json
{
  "created": "Mon, 31 Oct 2016 07:15:46 GMT",
  "mementos": [
    { 
      "memento-datetime": "Mon, 31 Oct 2016 07:02:28 GMT",
      "creator": "archive.org",
      "hash": "md5:2ba16e42ef ... sha256:fad5ec77bacc37495 ...",
    },
    { 
      "uri-m": "http://archive.is/PKqZl",
      "memento-datetime": "Mon, 31 Oct 2016 07:02:55 GMT",
      "creator": "archive.is",
      "hash": "md5:6425034d09 ... sha256:a541ad5ec70c05c70 ...",
    },
    { 
      "uri-m": "http://www.webcitation.org/6lfOC16B7",
      "memento-datetime": "Mon, 31 Oct 2016 07:03:40 GMT",
      "creator": "webcitation.org",
      "hash": "md5:97263c50c1 ... sha256:f89d62d1b7df28881 ...",
    }
  ]
}
```
Archive the Manifest
“You can’t tell the players without a scorecard” – Harry M. Stevens
Verifying the Authenticity of a Memento

• Given a Memento, URI-M, that we wish to verify
• Lookup the URI-M at a manifest server
  o e.g, captureproject.org/{URI-M}
• Discover all the mementos of the manifest, and verify their integrity with “trusty URIs”
• For each URI-M listed in the manifest, repeat the fixity calculation as described in the manifest
• Vote if fixity matches (not tampered with) or if fixity doesn’t match (tampered with)
  o Majority vote wins (assuming independent archives)

Video at https://www.youtube.com/watch?v=EY15lj-7_lc
Discussion
To the Rescue of the Orphans of Scholarly Communication

Herbert Van de Sompel
Los Alamos National Laboratory @hvdsomp
http://orcid.org/0000-0002-0715-6126

Michael L. Nelson
Old Dominion University @phonedude_mln
http://orcid.org/0000-0003-3749-8116

Martin Klein
Los Alamos National Laboratory @mart1nkle1n
http://orcid.org/0000-0003-0130-2097

The project is funded by the Andrew W. Mellon Foundation

@hvdsomp, @phonedude_mln, @mart1nkle1n
CNI Spring 2017, Albuquerque, NM, 3 Apr 2017