A Web-Centric Pipeline for Archiving Scholarly Artifacts

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

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

The Scholarly Orphans project is funded by the Andrew W. Mellon Foundation
Scholarly Orphans – Project Motivation
Research and Research Communication on the Web

- Consideration
  - Researchers are increasingly using a variety of web platforms for collaboration and communication

- Why?
  - Many of these platforms have desirable characteristics
    - Versioning
    - Time stamping
    - Social embedding
  - Their institutions do not provide platforms that have global reach
    - Collaboration, cf. Github ~ productivity
    - Communication, cf. SlideShare ~ visibility
Emma Schymanski

https://orcid.org/0000-0001-6868-8145

https://github.com/schymane
https://www.slideshare.net/EmmaSchymanski
https://figshare.com/authors/Emma_Schymanski/5087039
https://publons.com/author/1538491/emma-schymanski#profile
Shawn Jones

https://orcid.org/0000-0002-4372-870X

http://www.shawnmjones.org/
https://github.com/shawnmjones
https://www.slideshare.net/shawnmjones
https://www.blogger.com/profile/17827543974149663194
Research and Research Communication on the Web

- Consideration
  - Researchers deposit artifacts in these web platforms

- Web Platforms:
  - Dedicated to scholarship:
    - Commercial: e.g., FigShare, Publons
    - Not for profit: e.g., OSF, Zenodo
  - General purpose:
    - Commercial: e.g., GitHub, SlideShare
    - Not for profit: e.g., Wikipedia, Wikidata
Research and Research Communication on the Web

• Consideration
  • *Researchers deposit artifacts in these web platforms*

• Status quo - The researchers’ institutions commonly:
  • Do not know about the existence of these artifacts
  • Do not have a copy of these artifacts
Research and Research Communication on the Web

• Consideration
  • *Researchers deposit artifacts in these web platforms*

• Status quo – Uncertainty regarding long-term accessibility of these artifacts:
  • General purpose platforms don’t provide long-term access guarantees; platforms dedicated to scholarship commonly do
  • Uncertainty regarding the sustainability of unhindered long-term access to artifacts in these platforms:
    • Commercial: when is the change in business model coming?
    • Not for profit: will the next round of grant applications, member contributions be successful?
Research and Research Communication on the Web

- Consideration
  - Researchers deposit artifacts in these web platforms

- Status quo - These artifacts are not systematically archived:
  - No frameworks like LOCKSS/Portico exist for these artifacts
  - Researchers only selectively deposit artifacts in portals that provide archival guarantees; to obtain a cite-able DOI
  - Can’t expect researchers to (also) upload all artifacts in IRs
  - Web archives only incidentally archive these artifacts
    - Anecdotal & Hiberlink evidence
Emma’s SlideShare Artifact: 0 Mementos

https://www.slideshare.net/EmmaSchymanski/dmcm2018-community-resources-connecting-chemistry-and-toxicity-knowledge
http://timetravel.mementoweb.org/
Shawn’s GitHub Artifact: 1 Memento

https://github.com/shawnmjonesmediawiki
http://web.archive.org/
Web resources referenced in Elsevier corpus (1996-2012) without representative Memento in public web archives

https://doi.org/10.1371/journal.pone.0115253
The Need for an Archiving Infrastructure

Herbert Van de Sompel & Andrew Treloar (2014) A Perspective on Archiving the Scholarly Web
## Recording versus Archiving

<table>
<thead>
<tr>
<th><strong>Recording</strong></th>
<th><strong>Archiving</strong></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
Scholarly Orphans – Project Overview
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 (i.e., the scholarly artifacts deposited in web portals) for long-term archiving?

• Experimental project, aimed at exploring technical possibilities
The Scholarly Orphans Project

• Explores an institution-driven paradigm

• Academic institutions typically have a long shelf life
  • A basic premise underlying e.g., LOCKSS, perma.cc

• An academic institution should be interested in capturing the artifacts (intellectual property) its scholars deposit on the web

• Collecting and archiving such artifacts aligns with the mission of academic libraries
An Institutional Perspective

candidates for capture

Platform 1
Artifact Q

Platform 2
Artifact R
Artifact S

identity a
identity b
identity c
The Scholarly Orphans Project

• Explores a paradigm inspired by web archiving
  
  • Scale of the problem
  
  • Can’t expect researchers to upload all artifacts in an institutional repository
  
  • Bilateral agreements for archival purposes with most web portals unlikely
A Web Archiving Perspective
Inspiration

• LOCKSS
  • Web crawling approach
  • Focused on journal literature

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

• Institutional repository, auto-discovery of journal articles
  • Capture an institution’s output
  • Focused on journal literature

• The Locker Project & Amy Guy’s Personal Web Observatory work
  • Capture an individual’s web presence
  • Not focused on scholarly orphans
Scholarly Orphans – Prototype Pipeline Overview
Prototype Pipeline

- Track Artifacts
- Capture Artifacts
- Archive Artifacts

Pipeline Orchestration

Event Database
Demo - myresearch.institute

```
myresearch.institute

Track Artifacts ----> Capture Artifacts ----> Archive Artifacts

Pipeline Orchestration

Event Database

institutional processes
institutional event database
institutional archive
cross-institutional tools

cross-institutional processes
institutional event database
cross-institutional web archive
cross-institutional tools
```
myresearch.institute - Researchers

- Uniquely identified by ORCIDs
- Web identities in multiple portals
- Create various types of artifacts
myresearch.institute - Portals

- Tracking started August 27 2018
- Tracking artifacts created starting August 1 2018
- >2,200 artifacts tracked to date for all 16 researchers
myresearch.institute - Artifacts

• schema.org typology:
  • Answer
  • Article
  • BlogPosting
  • Comment
  • Dataset
  • PresentationDigitalDocument
  • Question
  • Review
  • SoftwareSourceCode
Tracking Artifacts
Tracking Artifacts - Description

• In order to track artifacts that were recently deposited by an institutional researcher in a portal, one reasonably needs:
  
  • The web identity of the researcher in the portal
    • Algorithmic discovery
    • Discovery via a registry
Algorithmic Discovery of Web Identities

From PostDoc database
Not up to date

Per person information elements

Web

Network identities
Updated information elements
New information elements
Social network connections

Per person information elements

App 1
App 2
App 3
App 4

## Discovery of Web Identities via a Registry (ORCID)

<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>

Martin Klein and Herbert Van de Sompel (2017)
Shawn’s ORCID Record

https://orcid.org/0000-0002-4372-870X
Emma’s ORCID Record

https://orcid.org/0000-0001-6868-8145
Tracking Artifacts - Description

• In order to track artifacts that were recently deposited by an institutional researcher in a portal, one reasonably needs:

  • The web identity of the researcher in the portal
    • Algorithmic discovery
    • Discovery via a registry

  • A portal API that supports:
    • Access by web identity
    • Access to contributions “since …” for the web identity

• Result of tracking:
  • URI(s) of new artifact(s) discovered in the portal
Tracking Artifacts - Architecture
Tracking Artifacts - Implementation

- Tracker event notifications:
  - Linked Data Notifications (JSON-LD) using AS2, PROV-O, schema.org
  - Identifiers: Unique tracker event identifier per notification
  - Dates: artifact publication date & artifact tracked date
  - URIs: 1+ artifact URI

- Event database:
  - Notifications stored indexed in ElasticSearch

- Researcher database:
  - SQLite
Tracking Artifacts - Demo

Demo: https://myresearch.institute/
Tracking Artifacts - Challenges

- Discovery of web identities of researchers
  - Algorithmic, registry-based currently not adequate
  - Fallback: manual discovery and entry
    - With help of researcher

- Portal API access by web identity
  - Broadly supported by general purpose portals
  - Typically not supported by scholarly portals
    - Some lack an API altogether
    - Should add ORCID access to APIs
    - OAI-PMH and ResourceSync need sets per web identity

- Professional versus personal contributions

- Tracking frequency/scale
Capturing Artifacts

Diagram:

- Track Artifacts
- Capture Artifacts
- Archive Artifacts
- Pipeline Orchestration
- Event Database
Capturing Artifacts - Description

• The capture process takes as input the URI of a new artifact discovered in a portal

• Its task is to create a representative institutional capture of the artifact

• Result of capture:
  • WARC file for new artifact in an institutional archive
Capturing Artifacts - Description

• Challenges:
  • Delineate the web boundary of the artifact
    • More than the input artifact URI
    • The boundary is in the eye of the beholder
  • Create a high-fidelity capture using an approach that scales for a steady stream of new artifacts
    • Unsolved problem
Capturing Artifacts

the page at URI-A is not the artifact

the artifact includes resources reachable via URI-A
Capturing Artifacts

low-fidelity memento via Internet Archive

WayBackMachine

low-fidelity memento via Internet Archive

capture approach designed for scale, not fidelity
not acceptable for representative capture of artifact

Scholarly Communication: Deconstruct and Decentralize?
A Bold Speculation Featuring

Alice  Bob  Carol

Herbert Van de Sompel
@hvdsomp
Los Alamos National Laboratory

Paul Evan Peters Lecture

@mart1nkle1n @hvdsomp
TPDL2018, Porto, Portugal, 12 Sep 2018
Capturing Artifacts

capture approach designed for fidelity, not scale
manual navigation of resources reachable via URI-A for all artifacts not feasible
Memento Tracer - Framework

http://tracer.mementoweb.org
Capturing Artifacts - Architecture
Capturing Artifacts - Implementation

• Capture event notifications:
  • Identifiers: Unique capture event identifier per notification; preceding tracker event identifier conveyed as provenance
  • Dates: Datetime of WARC file creation
  • URIs: 1+ WARC file URI

• Tracer, client-side:
  • Tracer Chrome extension leveraging Selenium IDE

• Tracer, server-side:
  • Stormcrawler; Selenium (Chrome) with Tracer plug-in; WarcProxy; file-system storage for WARC files

http://stormcrawler.net/
https://www.seleniumhq.org/projects/webdriver/
https://github.com/odie5533/WarcProxy
Capturing Artifacts - Demo

Demo: https://myresearch.institute/
Capturing Artifacts - Challenges

• Memento Tracer:
  • Language used to express Traces (interoperability)
  • Organization of the shared repository for Traces
  • Limitations of the browser event listener approach for recording Traces
  • Selection of a Trace for capturing a web publication by other means than URI pattern

• Legal constraints
Archiving Artifacts

- Track Artifacts
- Capture Artifacts
- Archive Artifacts

Pipeline Orchestration

Event Database
Archiving Artifacts - Description

- The archiving process takes as input the URI of a WARC file generated by the capture process.
- Its task is to ingest the WARC file in a cross-institutional web archive.
- This can be achieved using off-the-shelf web archiving software, e.g., pywb, Open Wayback.
- Result of archiving:
  - Mementos pertaining to newly discovered artifact in a cross-institutional, Memento-compliant web archive.
Archiving Artifacts - Architecture
Archiving Artifacts - Implementation

- Archiver event notifications:
  - Identifiers: Unique archiver event identifier per notification; preceding tracker/capturer event identifiers conveyed as provenance
  - Dates: WARC file ingest date; Memento-Datetime values
  - URIs: 1+ Memento URI, each corresponding to an artifact URI

- Web Archive:
  - pywb

- Social card:
  - MementoEmbed

https://github.com/webrecorder/pywb
https://github.com/oduwsdl/MementoEmbed
Archiving Artifacts - Demo

Demo: https://myresearch.institute/
Archiving Artifacts - Challenges

• Attempted to use ipwb, a pywb version that uses IPFS
  • Cross-institutional distributed file system with redundancy
  • Ran out of time to get it operationally stable

https://doi.org/10.1145/2910896.2925467
Scholarly Orphans – Summary
Summary (1/2)

• The Scholarly Orphans project explores an institution-driven approach to capture scholarly artifacts deposited in web portals

  • Artifacts out of scope of existing archival approaches such as LOCKSS, Portico, web archives

  • Institutions have a long shelf life, should be interested in collecting these artifacts, and have feasible scale for identity/artifact discovery
Summary (2/2)

- Components of the experimental pipeline:
  - Tracker: Automatically discover artifacts because researchers will not upload them to the institution
  - Capturer: High fidelity artifact captures through crowd-sourcing navigation patterns with Memento Tracer
  - Archiver: Cross-institutional, Memento-compliant scholarly web archive
Acknowledgments

- Los Alamos National Laboratory:
  - Lyudmila Balakireva
  - Martin Klein
  - James Powell
  - Harihar Shankar
  - Herbert Van de Sompel

- Old Dominion University:
  - Sawood Alam
  - Grant Atkins
  - Shawn Jones
  - Mat Kelly
  - Michael L. Nelson

- myresearch.institute – all volunteering researchers
A Web-Centric Pipeline for Archiving Scholarly Artifacts

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

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

The Scholarly Orphans project is funded by the Andrew W. Mellon Foundation

@mart1nkle1n @hvdsomp
TPDL2018, Porto, Portugal, 12 Sep 2018