If These Crawls Could Talk: Studying and Documenting Web Archives Provenance

Emily Maemura, PhD Candidate
Faculty of Information, University of Toronto
NetLab Forum Forum February 27, 2018
The Team

Nich Worby
Librarian, Web Archivist

Christoph Becker
Assistant Professor
Director, Digital Curation Institute

Emily Maemura
Doctoral Candidate

brought together through the Digital Curation Institute
McLuhan Centenary Fellowship in Digital Sustainability 2016-17
and supported by SSHRC

Ian Milligan
Associate Professor, Digital Historian

UNIVERSITY OF TORONTO LIBRARIES
UNIVERSITY OF TORONTO FACULTY OF INFORMATION
Digital Curation Institute
UNIVERSITY OF WATERLOO FACULTY OF ARTS
Department of History

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada
Pages from “policyalternatives.ca”

Sept. 2007 - Nov. 2009
over 200,000 pages per crawl

less than 50,000 pages per crawl
How are web archives made and used?
How can we document or communicate this?

Creating web archives

As a web archivist...
What do I need to document for researchers using web archives material?

Using web archives

As a researcher...
What do I need to ask about this data to have confidence in the analysis and findings?
How are web archives used by researchers? Can shared conceptual frameworks of research methods used with Web Archives collections help to systematize practices, advance the field, and make it easier to introduce new researchers to the area?

http://hdl.handle.net/1807/74866
“Research Objects” in Computational Science

Bechhofer et al., 2013, Why Linked Data is Not Enough for Scientists, Future Generation Computer Systems 29(2), February 2013, Pages 599-611

www.researchobject.org
Our Approach

Adopting the Research Object (RO) Framework for Research with Web Archives

- to structure the complex aggregation of computational processes, services, forms of data, contexts, and approaches
- balancing systematic approaches from computational sciences with humanistic issues of provenance and trust

Developed here to characterize three cases of web archives research studies, examples completed by co-author Ian Milligan
“Research Objects” as a Conceptual Framework

ORGANIZATIONAL CONTEXT
Ethical and governance approvals, investigators, etc. Acknowledgements

QUESTIONS
state a problem and/or hypothesis

STUDY DESIGN
scope, rationale for choices in tools, sources, methodology

ANSWERS
publications and presentations

DATA
materials studied, those taken as inputs to processes

METHODS
the tools, workflows, scripts, processes, settings, configurations, used to perform the analysis

RESULTS
materials produced derived datasets, visualizations, etc.
Case: GeoCities Community

- sole-authored research by a historian
- no ethics approval since data is publicly available via Wayback Machine
- research agreement signed with Internet Archive limits sharing and publication of specific derivative datasets
Case: GeoCities Community

Research Question: Did GeoCities users have a sense of community?

- also understanding what tools and approaches historians need to study the web
- iterative approach, exploratory investigation of archived GeoCities.com data
- development of Warcbase analytics platform
Methods

Analysis Phase 1 (2013):
data from Archive Team torrent (wget geocities.com)

~1TB
mirrored GeoCities directories

Prepare Archive Team Data
- Extract URLs
- Extract plain text

Analysis with Mathematica scripts, bash (on sample)

RESULTS

Analysis Phase 2 (2016):
data from Internet Archive end-of-life crawl

~4TB
GeoCities WARC

Prepare WARC data
- Import to warcbase

RESULTS

network graph visualization
word clouds
images

Analysis of raw link structures
popular images ordered by MD5 hash
An Initial Profile for Web Archives Research Objects

Disciplinary perspectives of researchers, roles in large teams and partnerships

Legal agreements and contracts that impact data sharing and use

Motivations for the study and research contributions - for whom is the work relevant?

How was data sourced? Collection timeframe(s) formats and interoperability

Which scripts, services, software packages used - was code published?

Are results FAIR, published, citable? How were they validated?

Interpretation of results is an important part of humanities scholarship, described in publication of findings (not simply the results of workflows)

Designs vary widely by discipline and conceptual perspectives

Include rationale for selecting sources, methods, scope
# Template and Workshop at RESAW 2017, London

## ORGANIZATIONAL CONTEXT

<table>
<thead>
<tr>
<th>Discipline(s) of Research Team</th>
<th>Funding</th>
<th>Approvals</th>
<th>Partnerships</th>
<th>Agreements &amp; Contracts</th>
<th>Acknowledgements</th>
</tr>
</thead>
</table>

## QUESTIONS

- Research questions and motivations for the study

## STUDY DESIGN

- **Scope**
  - limits of time period, geographic area, etc.
- **Rationale**
  - for choices of tools, sources, methods

## ANSWERS

- Publications and Presentations of findings

## DATA, METHODS & RESULTS

- **Data Sources**
  - metadata, method of collection
- **Data Preparation**
  - workflows and derived datasets
- **Data Analysis**
  - methods, config settings, logs
- **Results Generated**
  - published figures, data, code
**Template and Workshop at RESAW 2017, London**

<table>
<thead>
<tr>
<th>Questions</th>
<th>Study Design</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research questions and motivations for the study</td>
<td>Scope limits of time period, geographic area, etc.</td>
<td>Publications and Presentations of findings</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organizational Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discipline(s) of Research Team</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data, Methods &amp; Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Sources</td>
</tr>
<tr>
<td>metadata, method of collection</td>
</tr>
</tbody>
</table>

**Where do these come from, how are they made?**
Web Archivist Perspective: “Elements of Provenance”


How are web archives made? Which individual decisions are made in web archives practices, and how can we understand and communicate the impacts on resulting collections?
Our Approach

Starting with *Web Archiving Life Cycle Model* (Bragg et al. 2013) as a framework:

- What decisions are at each life cycle phase? (e.g. Appraisal and selection; Scoping; Data Capture; Storage and Organization; Quality Assurance and Analysis)
- Studying the process of using Archive-It to create three web archives collections by University of Toronto Libraries (UTL)
# Overview of the Collections Studied

<table>
<thead>
<tr>
<th>Collection / Timeframe</th>
<th>Canadian Political Parties</th>
<th>Pan Am Games</th>
<th>Global Summitry</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collection Timeframe</strong></td>
<td>October 2005 to present (ongoing)</td>
<td>February 2015 to December 2016</td>
<td>June 2016 to present (ongoing)</td>
</tr>
<tr>
<td><strong>Crawl frequency</strong></td>
<td>Quarterly (every 3 months)</td>
<td>Combination of Daily, Weekly, Monthly and One-time crawls</td>
<td>TBD, based on timing of summit events</td>
</tr>
<tr>
<td><strong>Crawl duration</strong></td>
<td>3 days</td>
<td>Varies widely by crawl (from hours to days)</td>
<td>5 days (currently test crawls only)</td>
</tr>
<tr>
<td><strong># of Active Seeds</strong></td>
<td>62</td>
<td>434</td>
<td>167</td>
</tr>
<tr>
<td><strong>Total data archived</strong></td>
<td>&gt;900 GB &gt;29,000,000 documents</td>
<td>&gt;100 GB &gt;3,500,000 documents</td>
<td>&gt;400 GB &gt;5,000,000 documents</td>
</tr>
<tr>
<td><strong>Crawl limits and rules specified</strong></td>
<td>Ignore robots.txt</td>
<td>Ignore robots.txt Block twitter.com URLs for “lang=?”</td>
<td>Ignore robots.txt</td>
</tr>
</tbody>
</table>
General Workflow for UTL crawls

Importance of iterations, different types of crawls: test crawls, production crawls, patch crawls

Managing organizational data budget across multiple collections
Three Key Findings

1) **Scoping:** decisions are made throughout

2) **Process:** Unforeseen issues arise during a crawl - the actions taken to resolve these issues need to be documented

3) **Context:** individual decisions interact and are influenced by changes in organizational context and wider environment, impacting the collection over time
Does the technical system allow?

☑ Yes, option available in Archive-It, since 2010

Does the legal environment permit?

☑ Yes, after Copyright Law Amendment, 2013

Does organizational policy guide action?

☑ Yes, new law interpreted in Permissions Policy, 2014

Individual curatorial choice to exclude sites with robots.txt for a particular collection or crawl
# Elements of Scoping, Process, Context to Document

<table>
<thead>
<tr>
<th>Element</th>
<th>Key Questions and Information to Document</th>
</tr>
</thead>
</table>
| **Motivation**           | What is the purpose of the collection?  
Has its mandate changed over time?                                                                                     |  |
| **Focus**                | Which geographic, temporal, technical, political, topical and/or social boundaries are defined to scope the collection?                                                                                                                      |  |
| **Access & Discovery**   | Who is the intended audience? Do they have known characteristics or needs?  
Which contractual, organizational, legal, or other agreements restrict access?  
What metadata fields and indexes support discovery? At what degree of granularity (by collection, site, or individual resource)?  
Which data formats or derivative datasets are available?                                                                 |  |
| **Seed list**            | What seeds were used in the scoping of the collection?  
What was the process of discovering and selecting seeds?                                                                                                                      |  |
| **Crawl timing**         | What is the frequency of crawls?  
How long do crawls run or what time limit is set?                                                                                                                           |  |
| **Crawl configuration**  | What settings control the depth of a crawl? For example, settings for capture by distance from original seed.  
Is the goal to have a more comprehensive or a breadth-focused collection?                                                                                                  |  |
| **Inclusions and Exclusions** | Are certain sites or media type included or excluded? For example, are regular expressions used to target certain files or directories in a URL structure? |  |
Next Steps - my work in Aarhus
Research Questions

How are web archives made and used *across different contexts*?

How can we document or communicate this *in systematic and comprehensive ways*?

+ How can this be supported by web archives research infrastructure?
Studying work with Netarkivet

Creating web archives

As a web archivist...
What do I need to document for researchers using web archives material?

Using web archives

As a researcher...
What do I need to ask about this data to have confidence in the analysis and findings?

Seeking specific projects and examples of connections between researchers and archivists at NetLab and the Royal Library
Approach and Methods

- **Initial Exploratory Work:**
  - Interviews and observation of work practices
  - Understanding workflows, creating diagrams, timelines
  - Identifying current needs and challenges

- **In-depth Analysis:**
  - More targeted discussions and interviews focused on particular data or artifacts
  - Digging into specific examples, connecting data traces to decisions and context of decision space
Anticipated Results and Outcomes

- Refine, revise, expand framework
- Report findings at NetLab Forum in May
- Compare and contrast with second case focused on work in Canada with web archiving and research (Summer/Fall 2018)
- Put it all together in a dissertation (!) “Logics of Order and Aggregation in Web Archiving Systems” (working title)
Thank You!

Emily Maemura, PhD Candidate
Faculty of Information, University of Toronto
e.maemura@mail.utoronto.ca

This research is supported by the Social Sciences and Humanities Research Council
Joseph-Armand Bombardier Canada Graduate Scholarship
and the Michael Smith Foreign Study Supplement