OCLC Member Forums
Fall 2016

Developing Best Practices for Web Archiving Metadata to Meet User Needs

oc.lc/wam

Jackie Dooley, Program Officer

… and other members of the OCLC Research Library Partnership Web Archiving Metadata Working Group

OCLC

OUR OBJECTIVE

Develop best practices for web archiving metadata that are community-neutral, output-neutral, and will

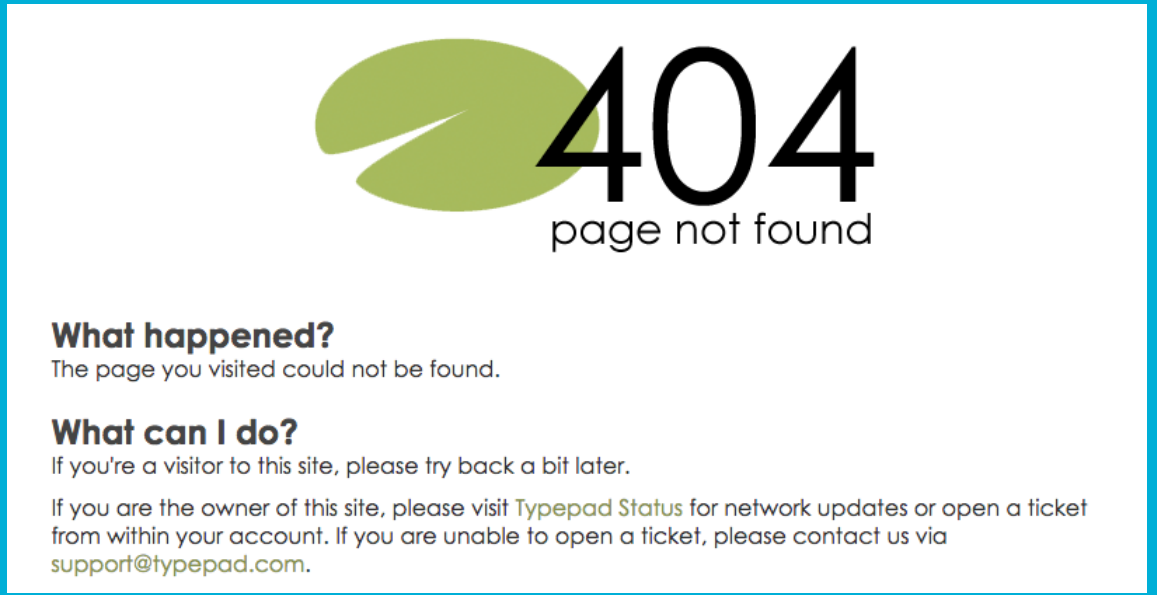
… consist of a set of data elements with content definitions

... not replace any existing standards

... be useable with any existing content or data structure standard

… interpret website characteristics in the context of metadata norms

THE IMPERATIVE



“Web archiving operates at the frontier of capturing and preserving our cultural and historical record.”

--The British Library web archive blog
14 September 2016

“In the past few years, we have noticed a significant uptick in the use of web archives in mainstream media.”

--Web Sciences and Digital Libraries Research Group
11 September 2016

“As hard as general preservation is, web preservation is even harder. Everything on the web dies faster.”

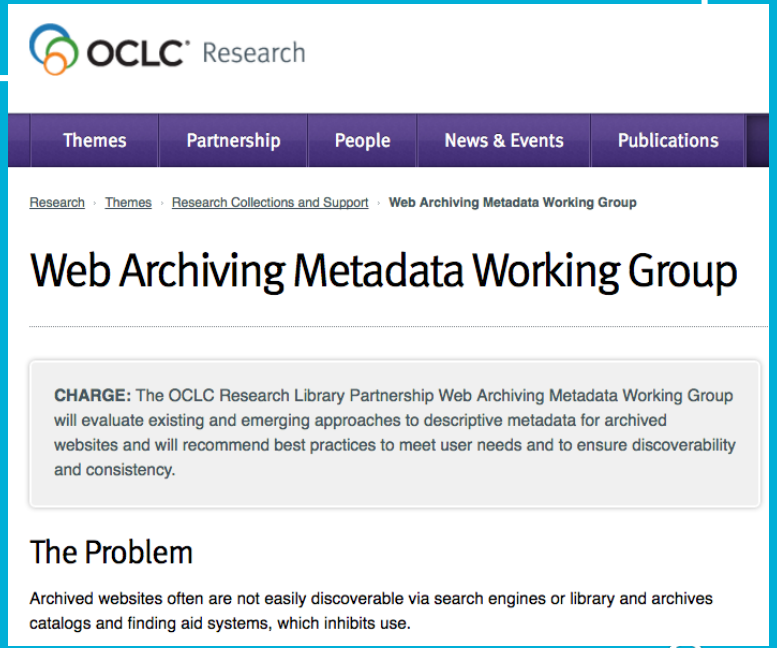
--Robin Davis. “Die hard: The impossible, absolutely essential task of saving the web for scholars” 2016

“It is far easier to find an example of a film from 1924 than a website from 1994.”

--M.S. Ankerson. “Writing web histories with an eye on the analog past” 2012

- Archived websites often are **not easily discoverable** via search engines or library and archives catalogs and finding aid systems, which **inhibits use**.
- **Absence of community best practices for descriptive metadata** was the most widely-shared web archiving challenging identified in two surveys:
  - OCLC Research Library Partnership (2015)
  - Rutgers/Weber study of users of archived website (2016)

OCLC RESEARCH LIBRARY PARTNERSHIP
WEB ARCHIVING METADATA WORKING GROUP



OCLC

Working Group charge

The OCLC Research Library Partnership Web Archiving Metadata Working Group will **evaluate existing and emerging approaches to descriptive metadata** for archived websites and will **recommend best practices to meet user needs** and to ensure discoverability and consistency.

oc.lc/wam

Planned outputs

- A report evaluating selected open-source **web archiving tools** will describe metadata-related functionalities.
- A report on **user needs and behaviors** will inform community-wide understanding of documented needs and behaviors as evidence to underlie the metadata best practices.
- **Best practices guidelines for descriptive metadata** will address aspects of bibliographic and archival approaches.

WHO IS IN TODAY'S AUDIENCE?

- Are you looking for a **basic or advanced** overview of web archiving issues?
- Is your institution **doing web archiving**?
  - If so, do you use Archive-It or other tools?
- Are you involved in **creating metadata** (for any type of material)?

CAN’T WE JUST USE RDA?

No! Must serve many communities

- We need output-neutral, community-neutral best practices
- Library catalogers are one of many metadata communities
- Item-level description of archived websites unfeasible at scale.
- Some RDA rules aren’t practical (Ex: edit a record whenever the site changes)
- Aspects of the archival approach to description may help meet users’ needs -- and vice versa

LITERATURE REVIEWS

Methodology

- Two literature reviews: **user needs** and **metadata**
- Selectively gathered published literature, conference reports and notes, social media, etc.
- Abstracted each item
- Synthesized our findings

Why study users' needs?

- A **necessary prelude** to development of metadata best practices
- We want to recommend an approach based on **real user needs & behaviors**
- **Users don't necessarily utilize the usual discovery tools** to locate archived websites
- **Lack of awareness** that libraries harvest and archive web content

We focused on three questions

- **Who** uses web archives?
- How and **why** do they use them?
- **What** can we do to support their needs?

Types of users identified in readings

- **Academic** researchers (social sciences, history)
- **Legal** researchers
- **Digital** humanists and data analysts
- Web and **computer** scientists

Note: We did not find sources that address other types of users.

Types of use

- **Reading** specific web pages/sites
- **Data** and text mining
- **Technology** development

What we learned about user needs

- **Formatting** and organization of data is an issue
- Lack of **discovery tools** make access challenging
- **“Provenance”** information is a critical missing piece
- Libraries and archives need to actively engage in **outreach** to users

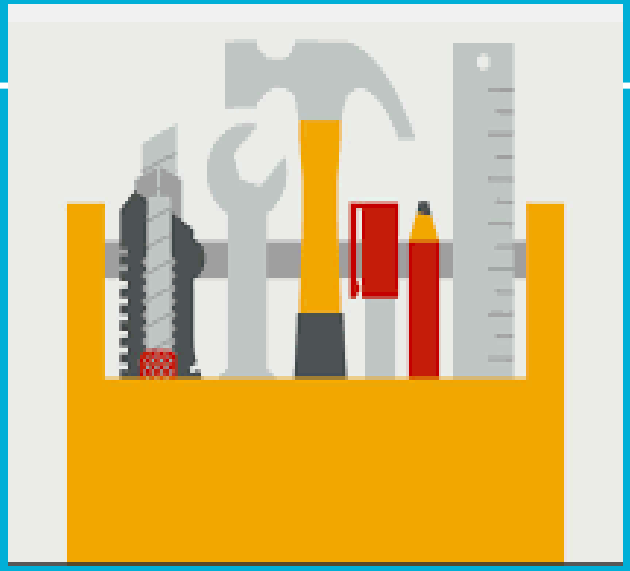
Topics in metadata readings

- Acquisition
- Archival approach
- Bibliographic approach
- Standards
- Tools
- Users
- Workflow

What we learned about metadata issues

- Confirmed that no specific set of best practices exists
- **Data formats** used include MARC, Dublin Core, MODS, EAD Finding aids
- **Metadata elements** used vary widely, and the same element may have different meanings
- **Level of description** varies as well: Single site, collection of sites, seed URLs …
- Creating metadata at **scale** is … impossible?

CAN WE AUTOMATE METADATA CREATION?



The community longs for the **magic bullet: automatic generation of metadata** from the tools used to crawl websites.

So we evaluated open-source tools

Archive-It
Heritrix
HTTrack
Memento
NetArchiveSuite
NutchWax
Site Story
Social Feed Manager
Web Archive Discovery
Wayback
webrecorder.io
WebCurator

Evaluation criteria

- Basic **purpose** of the tool
- Which **objects/files** can it take in and generate?
- Which **metadata profiles** does it record in?
- Which descriptive elements are automatically **generated**?
- Which descriptive elements can be **exported**?
- What **relation** does it have to other tools?

Evaluation grid for webrecorder

Criteria for evaluating tools	Description
Basic purpose of the tool and what it does. E.g. is it a capture tool (Heretrix) or display (Python Wayback) or an administrative layer (Archive-It) ...	Webrecorder is described as "a human-centered archival tool to create high-fidelity, interactive, contextual archives of social media and other dynamic content, such as embedded video and complex javascript." **Unlike other crawlers, Webrecorder archives web content through interactive browsing, capturing the exact sequence of navigation through a series of web pages or digital objects, and preserving the unique experience of an individual user** at a moment in time. This approach places control of the archive in the hands of the curator. The tool uses the **same software to capture and replay the site**; this approach is called symmetrical web archiving. The recording is **instantly archived as WARC files**. The files **can be replayed on the Webrecorder site or downloaded**. This is a **free service** that allows for **anonymous use or the ability to create public or private collections following registration** on a limited basis (100 MB

Etc.

What we learned

- Websites generally have **poor metadata** (e.g., title is “home page”)
- Tools have **few or no metadata-related features**
- Extractable data usually limited to title, crawl date ….
- Most tools capture only the site (usually in WARC format), so metadata **must be created manually/externally**
- A few tools enable manual input of metadata within WARC file
- Ergo: **no magic bullet** (big sigh)

DEVELOPING METADATA BEST PRACTICES

Why study existing practices?

- A **necessary prelude** to development of metadata best practices
- We want to **evaluate them in light of user needs**
- We want to **borrow their best features**
- We want to understand **how existing rules and guidelines are reflected** in current practice

Methodology

- **Analyze** descriptive metadata standards & local guidelines
- **Evaluate** existing records “in the wild”
- **Differentiate** bibliographic & archival approaches
- **Identify** issues specific to web archiving
- **Incorporate** findings from literature reviews

Descriptive standards under review

- *Describing Archives: A Content Standard* (SAA)
- *Integrating Resources: A Cataloging Manual* (Program for Cooperative Cataloging, based on RDA)
- Dublin Core

Local guidelines under review

- Archive-It
- Columbia
- Government Printing Office
- Harvard
- Library of Congress
- New York Art Resources Consortium (NYARC)
- University of Michigan
- University of Texas

Bibliographic vs. archival description

- Bibliographic: description of a single site
  - Based on RDA or Dublin Core
  - Includes standard bibliographic data elements
  - Usually does not provide context
- Archival: description of a collection of sites
  - Based on DACS or Dublin Core
  - Includes narrative description, e.g. of creator, content ...
  - Includes context of creation and use

Most common elements in local guidelines

Collection title	Language
*Creator/contributor	Publisher
Date of capture	Rights/Access conditions
Date of content	Subject
*Description	*Title
Genre	URL

* Only three elements that appear in all standards and local guidelines.

Some RDA elements that are not often used: Extent, Place of Publication, Publisher, Source of Description, Statement of Responsibility.

Analysis of data elements

- Definitions?
- Most frequently used?
- Core?
- Same concept, different elements?
- Same element, different concepts?

Analysis of records “in the wild”

- Data sources
  - WorldCat
  - ArchiveGrid
  - Archive-It
- Types of record
  - Bibliographic: MARC, Dublin Core, MODS …
  - Archival: MARC, finding aids

DILEMMAS SPECIFIC TO WEB CONTENT

1. Should the **title** be … transcribed verbatim from the head of the site? Edited to clarify the nature/scope of the site? Should it begin with "Website of the …"
2. Is the **website creator/owner** the … publisher? author? subject? all three?
3. Which **dates** are both important and feasible? Beginning/end of the site's existence? Date(s) of capture by the repository? Date of the content? Copyright?

4. How should **extent** be expressed? “1 archived website”? "1 online resource"? "6.25 Gb"? "circa 300 websites"?
5. Is the **institution** that harvests and hosts the site the … repository? creator? publisher? selector?
6. Does **provenance** refer to … the site owner? the repository that harvests and hosts the site? ways in which the site evolved?

7. Does **appraisal** mean … the reason the site warrants being archived? a collection of sites named by the repository? the parts of the site that were harvested?
8. Is it important to be clear that the resource **is a website**? If so, how to do so? In the extent? title? description?
9. Which **URLs** should be included? Seed? access? landing page?

Another issue: incomplete content

Physical Characteristics and Technical Requirements note:

"The web collection documents the publicly available content of the web page, it does not archive material that is password protected or blocked due to robot txt exclusions.

Although [institution] attempts to archive the entirety of a website, certain file types will not be captured dependent on how they are embedded in the site.

This can include videos (Youtube, Vimeo, or otherwise), pdfs (including Scribd or another pdf reader), rss feeds/plug-ins (including twitter), commenting platforms (disqus, facebook), presi, images, or anything that is not native to the site."

Source: New York University

MARC21 record types

When coded in the MARC 21 format, should a website be considered a …

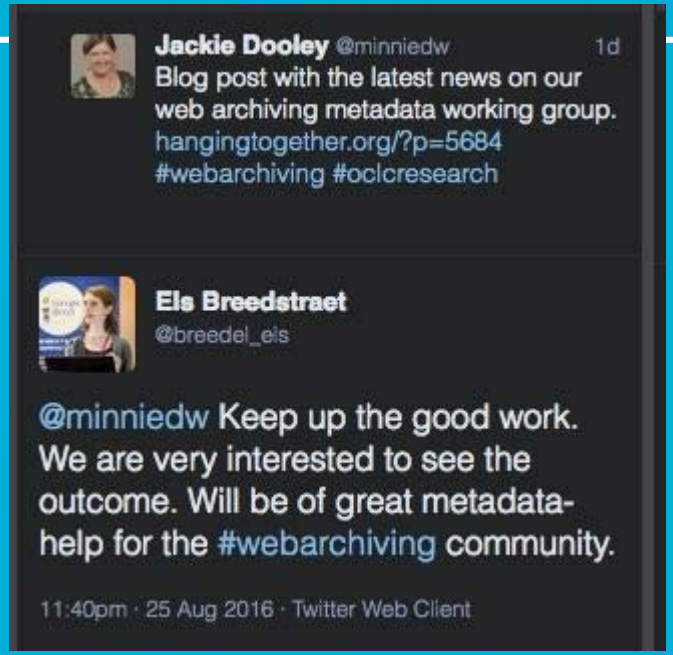
- Continuing resource?
- Integrating resource?
- Electronic resource?
- Textual publication?
- Mixed material?
- Manuscript?

FORTHCOMING REPORTS

Estimated publication dates

- **Tools** evaluation (January)
  - With evaluation grids
- **User needs** (February)
  - With annotated bibliography
- Best practice **guidelines** (April/May)
  - With annotated bibliography and local guidelines evaluation grid

DISCUSS!!


Jackie Dooley @minniedw 1d
Blog post with the latest news on our web archiving metadata working group. hangingtogether.org/?p=5684 #webarchiving #oclcresearch
Els Breedstraet
@breedel_els
@minniedw Keep up the good work. We are very interested to see the outcome. Will be of great metadata-help for the #webarchiving community.
11:40pm · 25 Aug 2016 · Twitter Web Client

OCLC Member Forums
Fall 2016

For more information, please contact:

Jackie Dooley
Program Officer, OCLC Research
dooleyj@oclc.org
@minniedw

oc.lc/wam

Because what is known must be shared.[SM]