Metadata for Web Archived Resources: Recommendations for Further Exploration

Rebecca Guenther

Oct. 30, 2015

1. INTRODUCTION

As part of the metadata consultancy for "Making the Black Hole Gray: Implementing the Web Archiving of Specialist Art Resources", I worked with NYARC staff to consider workflow issues related to the creation of catalog descriptions for archived websites and how that fit into the existing procedures and systems. I developed the *Metadata Application Profile and Data Dictionary for the Description of Websites with Archived Versions* that included detailed recommendations for elements to support discovery and use of these resources. Although the profile was written to support the existing environment in which catalogers integrated these procedures with other cataloging procedures, i.e. contributing records to OCLC using MARC, the profile has broader application as a data dictionary regardless of metadata scheme used. As part of that approach, mappings to other metadata structures were provided, with the goal of providing extensibility and interoperability.

During the development of the Metadata Application Profile, there was consultation with others in the community involved with the description of web archived resources. As a result, numerous questions arose about best practices and varying approaches. Although the profile provides a blueprint for describing these resources in the context of the NYARC environment, further investigation is needed to consider the issues that arose in developing a common method for describing web archived resources, reconcile different approaches, and look to future developments that will result in making the procedures and metadata specification sustainable and extensible.

This report, issued at the conclusion of the grant consultancy, establishes six areas of remaining challenges that will affect the future development of metadata policies for web archived resources, poses open questions that should be addressed, and concludes with six recommendations for next steps by NYARC. Because of the general lack of precedent in describing web resources, there are more questions than answers, and more recommendations for further investigation than specifications for what should be done.

2. CHALLENGES TO BE FURTHER EXPLORED

2.1. INTEGRATING BIBLIOGRAPHIC AND ARCHIVAL TRADITIONS OF DESCRIPTION

In order to enhance discovery and management of Web archives, libraries and other memory institutions require richer descriptions than current archiving tools provide. Whether they choose to describe web archives according to archival or bibliographic traditions largely depends on their cataloging environment and institutional policies about what department is responsible for web archiving. There has been discussion about how to apply archival principles of context, original order, and provenance to Web archives and how EAD might be leveraged. On the other hand, some institutions choose to describe web archives using bibliographic description standards and traditions to create metadata records.

Deciding how to apply long-standing traditions of description to web archived resources is a challenge, and either method could be justified, since this new form of material has peculiarities not applicable to traditional collections. Libraries and archives have long had to deal with the different approaches and have tried to integrate them in their environments and workflows. Often the two methods become complimentary using existing metadata standards, leveraging the strength of each. For instance, a collection level record might be created in the MARC-based catalog with a link to an EAD finding aid that provides more detail on the arrangement and contents of the collection. There are methods using existing metadata standards to provide hierarchical descriptions from a bibliographic point of view that could repurpose an archival description. Often there are tools that enable a transformation of a description from one standard to another (e.g. EAD to MODS and vice versa).

More exploration is needed to understand the strengths of each descriptive approach and build best practices for description and integration of varying methods. This task will include an analysis of how the two principles of description and accepted metadata standards may apply to Web archived resources. With active encouragement from NYARC, OCLC has started the conversation and is forming a group to consider metadata guidelines that integrate bibliographic and archival traditions. Some questions to consider are:

- What are the factors that help determine how web resources may best be organized and described? Based on past practices with analog materials, how could we integrate the two possible approaches—archival or bibliographical—in our discovery systems?
- In looking at the different ways to aggregate web resources, how do we define a collection?
- How do the concepts of original order, provenance, and context apply to web archived resources? What levels of description are needed and in what situations?
- Do we consider web captures like accession records in the archival sense? How will that metadata be conveyed? Might we consider the use of PREMIS events to record this information?
- How should appraisal decisions be described in the context of web archives, in order to document features such as the scope of the crawl, parts of websites that were missing for technical reasons, or resources that were blocked?

After considering this issue, there may be a need to enhance the Metadata Application Profile written for NYARC. It would be useful to add mappings to EAD if institutions want to apply the metadata elements in an archival description. There should be coordination with any OCLC initiatives as appropriate.

2.2. CONSIDERATIONS FOR MOVING TO A LINKED DATA ENVIRONMENT

A growing number of institutions are experimenting with Linked Data to assess its ability to provide richer information and combine data from different systems and communities. One effort is the development of a replacement format for MARC 21 based on Linked Data principles under the Bibliographic Framework Transition Initiative, BIBFRAME. A model and vocabulary was developed, which is currently under revision based on open discussion, and some tools have been made available. Numerous institutions are experimenting with describing resources using this emerging standard and others are pursuing grants to begin pilot projects. As experimentation continues to grow, there will be an attempt to describe a wider variety of resource types, which will inevitably result in changes to the model and/or vocabulary. The *Metadata Application Profile and Data Dictionary for Websites with Archived Versions* was written in terms of creating MARC records, although mappings to various other formats have been provided, including BIBFRAME (using the vocabulary in version 1). Analysis is needed to better understand how to model web archived resources using a Linked Data approach and how to describe them using the BIBFRAME vocabulary. Questions to consider are the following:

- How do we model web archives in terms of the Linked Data and simplified FRBR model adopted by BIBFRAME? How do live sites relate to the archival versions of those sites? Should these be separate descriptions if understood using a Linked Data model? How will they be linked?
- What changes are needed to the BIBFRAME vocabulary to accommodate these materials? Note that a version 2 is currently under development and expected to be released in early 2016.
- What kind of experimentation might we want to be doing to move to this new environment? Which organizations will participate?

It would probably be beneficial to have experimentation both inside and outside the NYARC institutions and perhaps as part of any pilot projects that may be grant funded.

2.3. EXTENDING THE USE OF THE METADATA APPLICATION PROFILE

It was intended that NYARC's Metadata Application Profile would be applicable to institutions outside of NYARC and beyond the art community. Broader review and revisions to the profile may therefore be required for wider use. The mappings provided in the profile were intended to suggest how it might be used in other metadata structure standards. Further analysis of the mappings and experimentation in applying it to other forms of resources is desirable.

- How do we bring in other partners outside of the NYARC community to use the Metadata Application Profile for describing their web archives? Who might those partners be?
- What is the process for discussion of changes requested?
- How will the profile be revised and maintained? How will NYARC be involved?

2.4. UNDERSTANDING USER DISCOVERY NEEDS

User studies are needed to determine what websites people are looking for and how they are searching for them. Some websites may not be easily found using web search engines (especially Google) and may benefit from more precise fielded searching. It is also necessary to consider where users expect the metadata to be stored—in a library catalog versus in web archiving tools. NYARC's recent implementation of ExLibris' Primo demonstrates how a discovery layer brings metadata from different sources together, including catalog records from Arcade and archived websites from Archive-It. How users are searching in that system and the sorts of results they are getting may be instructive here.

OCLC and Archive-It have been key players in providing tools for capturing and creating/using metadata for archived websites. NYARC (and the art community in general) would benefit from working with OCLC and Archive-It to improve integration of their systems, particularly in terms of indexing.

Further analysis might be informed by developing a set of use cases for the sort of information people are looking for, with consideration for websites that are most at risk of disappearing, and where we will have to rely on archived versus live sites.

A new development affecting the discovery environment is the formation of the International Image Interoperability Framework (IIIF)[1], which aims to enhance access to and interoperability among the rich resources in image repositories. How sharing of metadata and images on the web and how people access them using this framework may affect policies and practices in web archiving; this is an area for further exploration.

Another factor to consider is how user discovery tools may change as metadata is made available as Linked Data. As more and more data is serialized in a Linked Data compatible format, users may want to combine metadata from different sources to enhance their searching. Content providers are increasingly providing schema.org descriptions within HTML in their websites. Notably, OCLC transforms much of the MARC record into schema.org and other Linked Data compatible vocabularies in the HTML headers of WorldCat records.

Some questions to consider:

- How do we better understand what users are looking for in web archived resources?
- How is the metadata used and where should we concentrate our efforts in terms of description to best support users' needs?
- Is there a need for educating our users further about the availability and use of web archives?
- As metadata is provided in the headers of HTML documents using the schema.org approach, is there a way to apply and use it for archived websites?
- How might further integration of OCLC and Archive-It tools affect cataloging policies and practices. Will more or less metadata be needed?

2.5. REPURPOSING EXISTING METADATA

[1] International Image Interoperability Framework: Making the world's image repositories interoperable and accessible. http://iiif.io/

The current metadata environment involves a multitude of metadata standards that vary in richness, audience, and purpose. Metadata standards exist for different user communities, and we are often reliant on mappings, which are either imprecise or nonexistent. There are various approaches for bringing together results from different metadata schemes, which, because of the variety of detail and granularity in descriptions, can create problems for retrieval. Metadata that has been created is a valuable resource, and it is desirable to reuse what already exists. Interoperability is a key consideration, and there have been many studies analyzing approaches. In terms of potential users, analysis is needed as to how various forms of metadata might interact for live and archived websites.

- How do we repurpose existing metadata in varying formats that have been used to describe these materials?
- Which are the target metadata schemes that have been used for the description of live and archived websites?
- Will the mappings in the Metadata Application Profile provide a way to enhance interoperability with common descriptive elements? Is there a need for the specific communities that developed them to review and/or revise these mappings?
- Will tools be developed (e.g. XSLT) to promote transformations?

2.6. CONSIDERATIONS FOR LONG TERM PRESERVATION OF WEB ARCHIVED RESOURCES

The Metadata Application Profile developed for NYARC supports description for discovery and use of archived websites but has minimal information to support long-term preservation. We are reliant on internal, external, and communal digital storage and preservation repositories to provide that function. However, given the rapid change in hardware and software for both capturing and rendering archived websites, a more comprehensive long term strategy is necessary. Websites are complex and are made up of multiple file formats, which are subject to change and obsolescence. Therefore, complex hardware and software environments are needed to use them as originally intended. We have already seen the problems in archiving certain types of file formats and their inability to render; we can’t be certain that web archiving tools will be able to assess risk and enable us to use them in the future. Also, relying on internal links within the websites may not be a sustainable strategy.

One element in the Metadata Application Profile that supports preservation in terms of digital provenance is the Preservation actions. Currently, it is used to record the time period of the capture and the institution that was responsible. This element might be used to provide further metadata about actions on web resources, including supporting any preservation strategies.

- What should we be thinking about in terms of long term preservation of web archived resources and the metadata needed to support it?
- What actions other than capture should be recorded?
- Is there a need to provide structural metadata for these resources that are made up of multiple files? Now we are reliant on internal links, but these can break and result in lost data.

- How might provenance of the capture be described, especially in terms of archival appraisal decisions that may affect the ability to fully preserve an archived website? (See also above under 2.1)
- Is there a need to work with the International Internet Preservation Coalition to develop the Metadata Application Profile further to support long term preservation metadata needs?
- How are current archiving tools addressing the problem?
- What level and richness of description is necessary to enable the eventual migration and/or emulation of digital files contained within a WARC?

It is desirable to assess how the *PREMIS Data Dictionary for Preservation Metadata* might be used for these resources. Version 3 in particular addresses the problems of changing hardware and software environments for complex resources.

3. RECOMMENDATIONS

3.1. Convene a working group of stakeholders from the archival, library and museum community to analyze the applicability of traditions of description. Develop best practices for when to use each approach and how they might be integrated.

3.2. Consider an appropriate model for description of live and archived websites in terms of the BIBFRAME model, especially for art resources that are being archived. Participate in pilot projects that are experimenting with creating BIBFRAME metadata. Evaluate where enhancements to BIBFRAME are needed.

3.3. Engage the wider metadata/cataloging community in reviewing and commenting on the Metadata Application Profile. Have discussions about how it will be revised and maintained over time especially if its users are outside of the NYARC community.

3.4. Analyze user behavior in terms of discovery of archived websites and how current standards and tools support it. Develop use cases for what people are looking for and revise the Application Profile as needed based on these analyses. Analyze the effectiveness of searches in providing sufficient metadata that support the use cases and consider how it might affect cataloging policies and procedures, especially with improved integration of metadata in discovery tools. Work with Archive-It to refine and improve the use of metadata.

3.5. Analyze the current metadata environment and how a variety of metadata descriptions in various schemes might interoperate. This will follow on the study of user behavior and discovery needs in 3.4. Evaluate the role of IIIF and how web archiving tools and the WARC format might interact with image resources made available in that context.

3.6. Investigate the potential and benefit of providing metadata to support long term preservation. This will include an analysis of the challenges of capturing and maintaining such metadata. Evaluate existing digital preservation repository tools that can provide functionality for preservation.