Title

Preservation Metadata for Complex Digital Objects. A Report of the ALCTS PARS Preservation Metadata Interest Group Meeting. American Library Association Annual Meeting, San Francisco, June 2015

Authors

Drew Krewer and Chelcie Juliet Rowell

Body Text

At the 2015 ALA Annual Conference in San Francisco, the Preservation Metadata Interest Group meeting focused on pragmatic preservation metadata approaches for two complex content types: web archives and interactive new media art. The meeting drew approximately 40 attendees.

The session opened with a brief business meeting, which consisted of electing a new co-chair for the Preservation Metadata Interest Group. Jennifer L. Mullins (Dartmouth College) was elected to serve as co-chair from July 2015–June 2017.

Following the election, co-chairs Chelcie Juliet Rowell (Wake Forest University) and Drew Krewer (University of Houston) introduced the speakers Jefferson Bailey (Internet Archive) and Jason Kovari (Cornell University). As Director of Web Archiving Programs, Bailey manages the web archiving and digital preservation programs of the Internet Archive, particularly its Archive-It service. In his role as both Head of Metadata Services and Web Archivist, Kovari works on a wide range of metadata projects and digital preservation issues at Cornell University. Not in attendance was Bailey's co-presenter Maria LaCalle, a web archivist at the Internet Archive who works closely with the Archive-It service.

As more institutions include web archives in their digital collections, creating preservation metadata to support the long term stewardship of these files is a newly emerging challenge. Archive-It, a web archiving service of the Internet Archive, provides tools that over 370 partner institutions to harvest, manage, and provide access to archived web content. In their presentation "Don't WARC Away: Preservation Metadata & Web Archives," Bailey and LaCalle provided an overview of the WARC preservation format within the context of Internet Archive's Archive-It program and how preservation metadata such as PREMIS interacts with WARC files.

The WARC (Web ARChive) file format was developed in 2009 as a preservation file format for web archives. The format combines "multiple digital resources into an aggregate archival file together with related information" (Bailey & LaCalle, 2015, "Don't WARC Away"). A WARC format file concatenates one or more WARC records, each consisting of a simple set of text headers and a block of digital content. A WARC record is written by a web crawler as it goes out to the web and captures many different kinds of files (HTML, image files, and much more). There are eight types of records that can be found within a WARC file:

- warcinfo: contains information about the files within the WARC
- response: contains the full http response

- resource: contains the directly retrieved resource without protocol
- request: full http request with headers
- metadata: provides granular information about the harvested resource
- revisit: record that a site was revisited but was not recaptured since the site had not changed
- conversion: records that have been normalized on the fly
- continuation: allows completion across segmentation

A WARC file contains all sorts of technical metadata recorded at the point of capture by the web crawler, but it does not encompass everything. The format is not intended to package information related to rights and permissions, descriptions, agents and events, file format identification, validation, or characterization.

Although the hosted solution of Archive-It is not intended to be a preservation storage solution, not many institutions are not locally ingesting WARC files. According to a 2015 survey of Archive-It partners, 80% of respondents do not currently store local copies of their WARC files (Bailey & LaCalle, 2015, "Don't WARC Away"). Consequently, the question of preservation for web archives has largely not been raised. Bailey and LaCalle articulated challenges posed by the design of the WARC format and the PREMIS standard to the administrative uses of preservation metadata. Despite being metadata-rich, the WARC format is built on several premises that complicate its use with PREMIS:

- The WARC is concatenated in nature; one WARC contains thousands of digital objects.
- The resources that comprise WARC files are often dispersed for storage; there's little local acquisition.
- Because of the lack of file format verification possibilities inherent to a WARC, MIME types are unreliable, and format obsolescence becomes an issue.
- The volume of data is considerable.
- Crawlers are inconsistent in their preservation actions.

By the same token, other assumptions of the PREMIS standard present additional challenges for WARC files. PREMIS asserts that each bitstream is an individual object, which itself may have multiple composition levels. If a single WARC file contains thousands of digital objects, characterizing each object and its composition levels starts to become burdensome. Bailey and LeCalle propose several more "lightweight" approaches that might help institutions to initiate preservation actions on their WARC files. Data redundancy can be favored over record-level metadata granularity, crawlers can be used to generate mimetype reports, and PREMIS with its myriad decomposition levels could be replaced by a more pragmatic expression of event, agent, and object characteristics. Web archives are large, complicated collections. Bailey and LaCalle suggested that collecting organizations would do well to think more about institutional policies than technological details.

Just as web archives have provided unique digital preservation challenges, so has interactive new media art. In his presentation "In the Service of Art: Metadata for Preservation of Interactive Digital Artworks," Jason Kovari of Cornell University detailed the metadata strategies that arose from the Preservation and Access Framework for Digital Art Objects (PAFDAO) project, a two-year NEH-funded project that was

undertaken by Cornell University Library to find the most effective ways to preserve born-digital new media art. Though vitally important to understanding the development of media art and aesthetics over the past two decades, these materials are at serious risk of degradation and are unreadable without obsolete computers and software.

The pieces of art at the core of this project are highly experiential in nature; they are single-user multimedia from the early 1990s that were designed for small screens. These complex art objects were originally distributed via CD-ROM or the web and relied upon an elaborate set of interdependencies in order to render for the viewer as the artist intended the work to be experienced (Kovari, 2015, "In the Service of Art").

Before preserving such complex digital objects, one must determine what characteristics are being preserved and what level of preservation is most appropriate. Would each artwork be preserved and described at the file-level, or at the disk-image level? Researchers working on the PAFDAO project distributed a survey to artists, librarians, curators, and scholars to determine what characteristics should be preserved. Results indicated that preservation at the disk image level was most appropriate. A major concern was that too much information about the constituent parts of the digital object might result in over-description, thus obfuscating the most essential metadata needed to piece together a suitable rendering environment.

Legacy bibliographic records provided a "first pass" technical and artistic understanding of the digital art object. Certain MARC fields shed light on operating system requirements and other basic information. This metadata, while not comprehensive, does give librarians a starting point for selecting an emulation environment, as well as an idea of what the art should look like once it has been successfully rendered.

DFXML was used to store technical metadata that was created from running command line utilities. The information created from these utilities was stitched together and formatted into DFXML using Python scripts.

The Guymager disk imaging tool within the BitCurator environment was used to capture digital art objects stored on fragile media and devices. In addition to imaging each disk, Guymager provided reports on the following:

- sector health
- details about the creation environment
- three checksums
- unreadable sectors
- and more

The text file resulting from Guymager was then massaged with scripts to produce PREMIS metadata. In addition to the information automated through scripts, significant properties such as classifications for rendering environments were captured manually. The PREMIS also contains information about conservation treatments such as derivative image creation and other "life events."

While the automated metadata is fairly sophisticated, other metadata had to be captured manually and either placed within the PREMIS or in an associated text file containing a narrative. All artworks are given one of four classifications: executable files, browser-based, virtual reality, or macromedia. These classifications inform the strategies one would need to render an artwork. Other information captured included emulator documentation (such as configuration and setup of emulation software and any issues faced while running the emulator) and sector documentation (such as unreadable sectors and the implications unreadability has on determining the most faithful copy).

Future steps for the PAFDAO project include the release of a white paper, a potential release of the scripts used to create the PREMIS metadata, and a retrospective consideration of preservation metadata in light of the release of PREMIS v. 3, which may have stronger support for the preservation of software.

At the heart of implementing preservation metadata for both of these complex content types — web archives and interactive new media art — is a question about what constitutes the object being preserved and described. Both web pages and new media art objects are complex digital objects that can be broken down into innumerable further digital objects. Both presentations at this meeting of the Preservation Metadata Interest Group suggest that pragmatic preservation metadata implementations often seek to describe resources at a higher level than individual files.

References

Bailey, J., & LaCalle, M. (2015, June 27). Don't WARC Away: Preservation Metadata & Web Archives. Presentation to the Preservation Metadata Interest Group at the American Library Association Annual Conference. San Francisco, California. Retrieved from ALA Connect: http://connect.ala.org/files/2015-06-27_ALCTS_PARS_PMIG_web_archives.pdf.

Kovari, J. (2015, June 27). In the Service of Art: Metadata for Preservation of Interactive Digital Artworks. Presentation to the Preservation Metadata Interest Group at the American Library Association Annual Conference. San Francisco, California. Retrieved from ALA Connect: http://connect.ala.org/files/2015-06-27_ALCTS_PARS_PMIG_digital_art_objects.pdf.