Sharing, Using and Re-using Format Assessments

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 01970 USA
+1-617-495-3724
andrea_goethals@harvard.edu

Kate Murray
Library of Congress
101 Independence Ave
Washington, DC 20540 USA
+1-202-707-4894
kmur@loc.gov

Michael Day
The British Library
96 Euston Road
London NW1 2DP
+44 330 333 1144 ext. 3364
Michael.Day@bl.uk

Kevin L. De Vorsey
National Archives and Records Administration
One Bowling Green, Room 450
New York, NY 10004 USA
+1-212-401-1631
Kevin.devorsey@nara.gov

Jay Gattuso
National Library of New Zealand Te Puna Mātauranga o Aotearoa
Cnr Molesworth & Aitken St
Wellington, New Zealand
+64 4 4743064
Jay.Gattuso@dia.govt.nz

Paul Wheatley
Digital Preservation Coalition
37 Tanner Row
York YO1 6WP
+44 0 1904 601871
paul@dpconline.org

ABSTRACT

Many cultural heritage institutions with digital collections have performed assessments of file formats to inform decision making for a wide range of activities. The development of digitization standards and transfer requirements for depositors, the selection of storage and access platforms, and preservation planning and the decisions to use emulation or migration as a preservation approach all benefit from the assessment of file formats for their appropriateness in these different contexts. This workshop will bring together several institutions who have created format assessments, together with institutions who already are or could potentially reuse this work to inform their own institutional policies and preservation planning. The workshop will start with short presentations to expose the assessment work that has been done, followed by a discussion of how they are being used, or could be used, and possibilities for more effectively sharing these resources across institutions for the benefit of the digital preservation community.

Keywords

File Format Assessments; Recommended File Formats; Preservation Formats

1. FORMAT ASSESSMENTS

Several cultural heritage institutions responsible for preserving digital content have generated what could be called format assessments. While the specific workflow, criteria and artifacts created have tended to be institution-specific, and have even varied over time within the same institutions; at a high level all of these format assessments could be defined as the detailed documentation of the properties of a format to gain insight into the format's suitability to fill a repository function, e.g. as an archival format, as a transfer format, as an access format. They are used to support decisions related to content that may come into the repository, or that is already under management in the repository.

The institutions writing format assessments create them for various reasons. Some create them to inform policy and related guidelines for their preservation repository, for example as the basis for guidelines for content creators, or to restrict the formats accepted into their repository. Other institutions create them to inform the broader digital preservation community, potentially as the basis for best or at least good practices. Still others create them to make decisions about formats to select for normalization or migration targets, or to identify formats that might be at risk of obsolescence within their repository. Because of the diversity of the reasons for format assessments, among institutions creating them the methods used and the results are divergent. A key difference is how a "format" is defined or scoped in the assessment. Some, like those written for the Library of Congress' Sustainability of Digital Formats Web site [1] are very granular - there are eight different variations of the JPEG2000 JP2 format described on the site. In contrast the assessments written by the British Library [2] are not as granular, e.g. there is a single assessment covering the JPEG2000 JP2 format. The assessments created by Harvard Library [3] are also less granular than the Library of Congress' but include associated assessments of metadata and tools specific to formats.

Despite the differences in how and why format assessments are created, institutions of all types within the digital preservation community could benefit from the broader sharing of these assessments. There is a great deal of time, effort and resources that go into preparing format assessments. Leveraging the research and findings already done by other institutions will allow institutions to focus efforts on work not already done. In addition, institutions that do not have the resources to do their own format assessments are still able to write needed preservation policies and conduct preservation planning by reusing the work that has already been done. Lastly, exposing this work to more eyes in the community should lead to more discussion and constructive feedback about formats and their suitability for preservation that can lead to establishing community-accepted best practices in this area.

2. WORKSHOP STRUCTURE

The authors of this proposal represent institutions who are actively engaged in format assessments and wish to make this work discoverable and useful as a resource to the digital preservation community. The workshop will begin with brief descriptions from each of the authors on their role related to format assessments. The Library of Congress (LC) will present a case history perspective of its detailed PDF [4] assessments on the Sustainability of Digital Formats website to demonstrate the granularity of LC's assessments but also real life application of their usefulness. Harvard Library (HL) will describe the format and related metadata and tool assessments it has been doing as preparation for supporting new formats in HL's preservation repository. The British Library will outline the evolution of its file format assessment activities in the context of the development of a preservation planning capacity based on a deeper understanding of the Library's collections and preservation priorities. The National Archives and Records Administration (NARA) will describe the assessment

methodology it undertakes to determine file formats that are appropriate for use by Federal agencies transferring permanent electronic records. The National Library of New Zealand (NLNZ) will explore the relationship between format identification and format assessments. This will draw on a decade of experience and will discuss the boundaries between them, the institutional mandates that inform their use and value, and point to some directions on how this information is best linked and shared.

The presentations will be followed by an informal discussion to explore the following topics:

- Use of format assessments
    - Are they being used by institutions that did not create them, and for what purposes?
    - Is there any additional information or infrastructure needed to make them more discoverable, interpreted as intended by the creators, or more useful to other institutions?
- Central discovery platform for format assessments
    - Is it desirable for the community to have a central portal for format assessments, and if so, how would this be maintained?
    - What would be the challenges to implementing this and what are some ideas for resolving them?

The last part of the workshop will focus on identifying concrete next steps following the workshop, including but not limited to an update of this paper after the workshop.

3. REFERENCES

[1] Sustainability of Digital Formats: Planning for Library of Congress Collections http://www.digitalpreservation.gov/formats/

[2] File Formats Assessments – wiki.dpconline.org http://wiki.dpconline.org/index.php?title=File_Formats_Assessments

[3] Format Assessments – Harvard Library Digital Preservation – Harvard Wiki https://wiki.harvard.edu/confluence/display/digitalpreservation/Format+Assessments

[4] PDF (Portable Focument Format) Family http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml