The Evolving Process to Add Preservation Support for New Formats at Harvard Library

Andrea Goethals, Franziska Frey, David Ackerman; Harvard Library; Cambridge, MA, USA

Abstract

In 2000 Harvard Library began populating the Digital Repository Service (DRS), its digital preservation repository, with digitized text, images and audio. Over the years the Library continued to add preservation support to the DRS for new formats including born digital websites, PDF documents and email. Library collections however continued to grow and diversify to include a wide range of formats not supported by the DRS, including video and a variety of born-digital formats. In 2013 the Library began to bridge that preservation support gap by refining the process of how new formats are supported in the DRS. This paper describes the new process, which is a more consistent workflow and includes external expertise; as well as analysis tools that could be used by other institutions to broaden the range of digital formats that they are able to preserve.

Introduction

Over the last decade the collections at Harvard Library have diversified to include content in many different analog and digital formats. While preservation solutions exist for the more "traditional" analog and digital formats at Harvard, much of the media-based collections and born-digital material is at risk. The media collections, on VHS, mini-DV, U-matic video tapes, and many other carrier formats are at risk primarily because of media degradation and because much of the playback technology is now obsolete. The born-digital material is at similar risk of permanent loss for many reasons, but primarily because of format obsolescence, lost or corrupt bits, and lack of metadata and documentation needed to interpret the bits.

Curators and collection managers at Harvard have requested that many of these media and born-digital formats be supported by the Library's digital preservation repository, the Digital Repository Service (DRS). Support for new formats has been added to the DRS on an almost continuous basis since it was launched in 2000, but not fast enough to keep up with all of the demands (see Figure 1). To prevent permanent loss, the Library needed to speed up the process for adding support for new formats to the DRS.

This problem is not unique to Harvard. This is a reflection of the Information Age we live in where content now routinely comes in many different formats, from a variety of creators. If collecting institutions everywhere do not already have this problem, they will soon need to determine how to go beyond preserving well-understood digitized content to include the more complex media and born-digital formats.

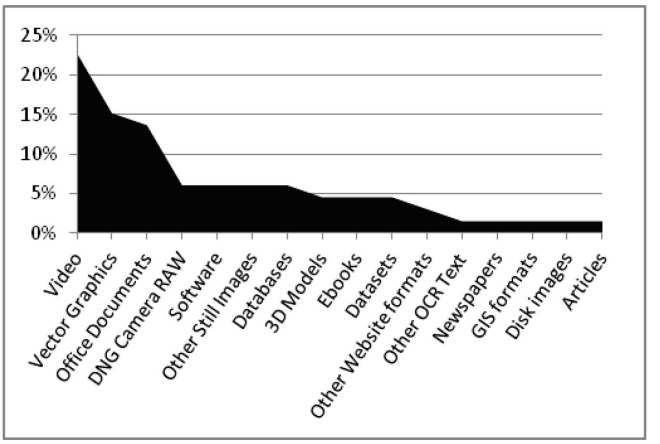

Figure 1. This graph shows the percentage of requests for DRS support for various formats made by Harvard curators and collection managers since 2004 for which there still was not preservation support in 2014. Almost a quarter of the requests have been for video formats, while there is a long tail of requests to support many other formats.

What is a Supported Format in the DRS?

All content preserved in the DRS conforms to one of twenty content models, for example Still Image, Audio, Document and Email Message. Each content model prescribes preferred and accepted formats, valid relationships between the object's files and to other objects, associated metadata schemas, delivery applications and associated preservation plans. Two content models allow files in any format - Opaque and Opaque Container. The main difference between the two is that the files of Opaque Container objects are aggregated into compressed zip files while the files of Opaque objects are left as separate files. Although these two content models allow for any file format to be deposited to the DRS, it is not considered that the DRS supports all formats because these files receive only bit-level preservation. Their disposition as opaque is considered to be temporary, until sometime in the future when the formats can be fully supported directly or as a result of a migration.

A format is "supported" by the DRS when preservation staff are reasonably confident that material in this format can be used now and that it can be made usable on an ongoing basis through preservation interventions. It requires policy and strategy decisions, the provision of guidelines and often tool development and enhancements. Supported formats can be deposited to the DRS, can be managed within a DRS administrative interface and can be delivered for use by researchers through Library-maintained APIs and delivery tools.

Specifically it requires the provision of specifications to collection managers and DRS depositors on the formats that are preferred and acceptable; content creation guidelines, for example, do not use encryption; and instructions on metadata or documentation that should accompany the content. All of this necessarily needs to be backed up by research and analysis.

It also requires that the DRS tools accurately understand the new format and support any technical metadata schemas or elements that are adopted for the format. The DRS deposit and management tools must accept and understand the format. Most importantly, discovery and delivery systems must exist for the format; otherwise the content cannot be used by researchers.

Before this project, it often took a great deal of time to add new formats to the DRS for several reasons. All of the analysis and development work was done by existing in-house preservation and IT staff who were also working on other projects and ongoing operations concurrently. Because the format support work was intermittent as time allowed, some time was lost to re-familiarization with the work, and the overall completion time was longer than if it had been uninterrupted. If preservation staff were unfamiliar with the formats or associated metadata, time would need to be spent on research as a first step before any other progress could be made. In addition, the overall process for adding new formats was somewhat ad-hoc and largely designed from scratch each time new formats were added.

New Approach

To make significant headway on adding preservation support for the most requested formats, a three-year project to fast-track support was launched with the support of the Arcadia Foundation. The formats were scoped to include video, office formats (word processing, spreadsheets and presentations), 2D vector formats including CAD, Adobe DNG raw camera images and 3D formats. A few formats that had not been formally requested were included for strategic reasons: disk images and image stacks. To support all these formats within three years the Library's format support process needed to be transformed into a routine that was much faster and could easily be repeated.

At a high-level the approach was to separate the preliminary analysis tasks from the software development work into two different sub-projects because they required different knowledge and skills. For any format, the analysis could be done as soon as expertise was available so that the specifications were already waiting when developer resources became available. In this way the analysis and implementation could be staggered in a way that was more efficient than trying to line up the staff resources needed for all of the analysis and development at one time.

Another significant change was that external consultants with specialized format expertise were sought to help preservation staff with the analysis. Besides temporarily increasing the pool of preservation expertise, it had several other benefits. If a consultant could be identified who already had the right format expertise, there was no lag time needed for the up-front format research. Also, having to specify to a consultant the analysis tasks and desired deliverables required preservation staff to clearly define the overall analysis framework in a way that could be repeated for other formats. And most importantly the analysis work for different formats could be done in parallel (provided that multiple different consultants were used), so that the elapsed time was shorter. As a result of these changes the overall methodology for adding support for new formats was made more consistent and efficient.

Results

An immediate result of this project is that the Library is able to make substantial progress in addressing the backlog of format support requests. By adding support for the video and born-digital formats described earlier, sixty-four percent of the preservation support requests will be addressed by the end of this three-year project. Many of these formats, such as CAD and 3D object models, are very complex and few if any institutions are confident preserving content in these formats [1]. The Library will share its format, metadata and tool decisions on its public website so that this information will be available to other institutions. In addition, because the workflow and analysis tools that were developed might also be of interest to other institutions, those are described here.

Workflow

At an operational level the Library now has a clear and repeatable workflow for adding support for new formats to the DRS. Table 1 lists the workflow's analysis and implementation tasks. In general the tasks are sequential but they are discrete enough that they can be distributed among internal staff and external consultants. The first task is to decide which tasks will be done internally or by consultants or by a combination, depending on in-house expertise and the ability to identify external experts. Table 2 shows how the analysis tasks were distributed among internal staff and consultants for each format group.

Table 1: Sequential steps of the new process to add support for formats to the DRS

Analysis (by preservation, format & metadata experts)	1. Divide up analysis responsibilities
	2. Format criteria (Determine key criteria to compare formats)
	3. Format analysis (Compare formats using criteria, identifying key pros and cons, recommend which to prefer & accept)
	4. Format profiles (Describe the subset that will be preferred and accepted)
	5. Preservation strategy (Short & long-term approach)
	6. Metadata analysis (Identify or develop technical and provenance metadata schemas)
	7. DRS content model
	8. Tool analysis (Determine tools to improve FITS and if applicable normalize formats)
Implementation (by software developers)	9. DRS deposit tools (Be able to identify &validate formats, extract metadata)
	10. DRS management tools
	11. DRS delivery tools

Table 2: The division of analysis tasks among Library staff and external consultants. Key: I = done by internal staff, E = done by external consultants, IE = done by a combination of internal staff and external consultants

Format group	Analysis Tasks (See Table 1 for Tasks)							
	1	2	3	4	5	6	7	8
Video	I	I	I	I	I	E	IE	E
Word processing	I	IE	E	E	IE	E	IE	E
2D vector	I	IE	E	E	IE	E	IE	E
3D formats	I	IE	E	E	IE	E	IE	E
DNG	I	IE	IE	E	IE	E	IE	E
Image stacks	I	IE	E	E	IE	E	IE	E
Disk images	I	IE	E	E	IE	E	IE	E

Format Assessment Criteria

Early in the project, Library preservation staff decided on the criteria to analyze and compare formats. The starting point was the criteria that had been used in a study by Ryan [2] on the factors leading to file format obsolescence. That list was slightly modified and then prioritized according to the degree to which staff thought the criteria could suggest the ease or difficulty of maintaining long-term access to files in a particular format. This prioritization of criteria, shown in Table 3, decreased the time needed for analysis as the criteria rated somewhat or not at all important could be largely ignored.

Table 3: The rating of criteria for comparing file formats

Importance	Criteria
Very important	• Cost to maintain environment for access or processing • Dependency on a single organization or company • Format expertise available • Legal restrictions affecting use now or long-term • Dependencies on particular SW/HW • Quantity & availability of rendering SW • Availability of specifications • Widespread use by consumers • Widespread use by professionals
Somewhat important	• Backward compatibility • Level of format complexity • Degree to which compression is understood • Ease of accurate validation • Degree to which specification is complete and understandable • Standardized • Storage requirements relative to other similar formats • Support for technical protection mechanisms • Archival use
Not very important	• Browser support • Community / 3rd party support • Descriptive metadata support • Developer / corporate support • Error-tolerance • Geographic spread • Lifetime • Revision rate • Technical metadata support • Malware

In addition to these generic criteria that could be used to analyze all formats, additional format-specific criteria were added for each format group. These criteria focused on key features or characteristics of file formats. Note that this is different from the concept of significant properties which usually refer to the properties of specific digital objects as in Andrew Wilson's well-accepted definition [3]. These criteria were added to help determine the suitability of a format as the archival master format (as opposed to a use copy), and to determine if key features would be lost in a conversion to the format. As an example, the format-specific criteria added for the video format analysis are:

- Ability to encode in true lossless compression
- Ability to encode in visually lossless compression
- Max chroma subsampling
- Max resolution
- Highest bit resolution
- Highest supported bitrate
- Compression ratio

Format Matrix

To analyze each format group, a format matrix was constructed. On one axis the formats under consideration were listed. Formats made the list if they met any of these conditions: they were known to be in Harvard collections, they were preferred or accepted formats for other preservation repositories, they were in common use either by the consumer market or the professional market, or they were 'emerging' formats in industry or in the standards community. The generic and format-specific criteria were placed along the other axis. The analysis consisted of doing research or experimentation to fill in the matrix cell values and then shading the cells using the common traffic light scheme where red indicates a high preservation risk, green indicates a preservation-friendly value and yellow indicates the values in-between. The result of the shading is that at a glance it becomes evident which formats are better or worse preservation format candidates. Out of this information two classes of formats were identified:

- Preferred formats: formats that will be encouraged and that will be used by Library reformatting labs
- Accepted formats: formats that are not preferable but are popular and well-supported currently, and/or there aren't equivalent formats; these may be normalized to another format before ingest into the repository

Format Profile Template

For use by the repository for ongoing preservation planning, the preferred and accepted formats are described using a template

developed for the project. The format profile includes the following sections:

- Full name, aliases, MIME media-type, file extensions
- Brief description
- Key adopters
- Original and current maintenance organizations
- Availability and location of specification
- Patent information
- Key related links
- Risk summary
- Mitigation of key risks
- References

These format profiles can be seen as simpler versions of the format descriptions maintained by the Library of Congress (LC) [4] in that the LC descriptions include much more detailed information about the formats and their relationship to other similar formats. The exception is that the profiles developed for this project explicitly include a description of the key risks and how they will be mitigated in the DRS. In other words they go beyond format description to include local preservation strategy.

Conclusions

On an organizational level, preservation staff now have experience with a new way of working with external experts to supplement internal expertise. The longer-term goal is that the Library will have a network of format experts that could be consulted at additional times during the lifetime of preserving the content, for example when content needs to be migrated.

As evident by the literature and current workshops [5][6][7]; there is a great deal of interest in establishing efficient workflows in core operational areas such as digitization, quality control, content description, and repository ingest. Less attention has been paid to refining workflows for ongoing repository maintenance, such as increasing the range of formats supported by the repository. This project shows that there are real benefits to streamlining these repository operations as well. At the time of this writing, analysis and implementation tasks are underway in parallel for seven of the format groups. This is only possible because four different consultant groups are currently helping with the work. If only internal staff were engaged in the work, only a small fraction of this work could be done.

The 2015 NDSA Strategic Agenda [8] concluded that it is impractical for every cultural heritage institution to develop their own in-house expertise in all areas, and suggested that a more practical approach is for institutions to specialize in some areas and rely on other institutions for other expertise. This project demonstrates an alternative approach, working with specialized external format consultants to supplement internal expertise. As shown in Table 2, the process developed in this project is collaborative; both internal and external staff participated in the analysis of each format group, taking on particular analysis tasks depending on the location of the expertise.

More experimentation and experience is needed to explore approaches to finding, leveraging and sustaining format expertise. It is clear that no one organization can rely completely on internal expertise, but it remains to be seen what the best approach will be. It may be that a combination strikes the best balance between immediate gains seen by incorporating external experts and the long-term benefits of establishing in-house experts.

References

[1] K. Rimkus, T. Padilla, T. Popp and G. Martin, Digital Preservation File Format Policies of ARL Member Libraries: An Analysis, D-Lib Magazine, 20, 3/4. (2014).

[2] H. Ryan, Who's Afraid of File Format Obsolescence? Evaluating File Format Endangerment Levels and Factors for the Creation of a File Format Endangerment Index, Dissertation, University of North Carolina Chapel Hill, (2014).

[3] A. Wilson, A Significant Properties Report, InSPECT Work Package 2.2, Draft/Version 2. (2007).

[4] Library of Congress, Sustainability of Digital Formats: Planning for Library of Congress Collections, Website http://www.digitalpreservation.gov/formats/index.shtml (Retrieved 2015)

[5] A. Neatrour, M. Brunsvik, S. Buckner, B. McBride and J. Myntti, The SIMP Tool: Facilitating Digital Library, Metadata, and Preservation Workflow at the University of Utah's J. Willard Marriott Library. D-Lib Magazine, 20, 7/8 (2014).

[6] B. LeFurgy, "Steps in a Digital Preservation Workflow," ALCTS Webcast. (2012).

[7] Indiana University, Workflows for Digital Preservation and Curation Workshop, Data to Insight Center. (2012).

[8] NDSA, 2015 National Agenda for Digital Stewardship. (2014).

Author Biography

Andrea Goethals is responsible for providing leadership in the development and operation of Harvard's digital preservation program and for the management and oversight of the Digital Repository Service (DRS), Harvard's large scale digital preservation repository. She leads the National Digital Stewardship Residency (NDSR) Boston program, participates in the International Internet Preservation Consortium (IIPC) Preservation Working Group, and is the co-chair of the National Digital Stewardship Alliance (NDSA) Standards and Practices Working Group.

Franziska Frey is the Malloy-Rabinowitz Preservation Librarian for the Harvard Library. She is responsible for shaping the strategic direction for preservation, conservation and digitization initiatives to ensure long-term access to all collections with the goal to create a seamless continuum for the long-term preservation of traditional collections and digital content across the Harvard Library. Before joining Harvard Library, she was the McGhee Distinguished Professor at Rochester Institute of Technology's School of Print Media and a faculty member in the Center for Imaging Science at RIT. She has been involved in several international standards groups, and currently serves as the chair of ISO JWG 26.

David Ackerman is the Head of Media Preservation for the Harvard Library. Prior to that he managed Audio Preservation Services for the Harvard College Library. He co-chairs AES SC-07-01, Working Group on Audio Metadata Standards and AES TC-ARDL, Technical Committee on Archives, Digital Libraries and Restoration.