

Steering Resources to Safe-Harbor Repositories: The Need for Reliable, Accurate and Affordable Ingest Services

*Stephen Chapman
Harvard University
Cambridge, Massachusetts, USA*

*Stephen L. Abrams
Harvard University
Cambridge, Massachusetts, USA*

Abstract

With the emergence of centralized, large-scale digital archives, geography becomes a key factor in the preservation of cultural heritage materials. Objects “inside” archives will be actively monitored and managed, whereas those “outside” will be at greater risk of loss and obsolescence. Developing ingest systems and services to process, package and transport objects into managed safe-harbor repositories is an immediate need.¹ Standards, frameworks, and business models for digital archiving must also evolve in due time to support these services.

Ingest solutions must address a range of challenges: legal, technical, and financial. Software development, however, is a logical starting point, since tools that automate pre-archiving tasks meet technical requirements for viability and economic ones for affordability. The key tasks to automate are production of preservation metadata, transformation and validation of formats, and creation of repository-compliant transfer packages.

The Harvard University Library (HUL) Office for Information Systems (OIS) has developed two applications to promote use of the HUL Digital Repository Service. JHOVE, developed with JSTOR, is a format-validation program; Dmart is a batch deposit tool for audio preservation packages. In Harvard’s experience, the target user for such applications has typically been a professional depositing agent with technical expertise, who consults as needed with curatorial experts.

With greater understanding of ingest requirements, and the profiles of persons or agencies likely to perform these services, it is hoped that industry will develop and support tools for the digital archiving market.

Introduction

Research libraries acquire, catalog, preserve and make available for use a wealth of information resources to diverse user communities. Artists, authors, audio engineers, photographers, and publishers, on the other hand, specialize in creating the resources that libraries acquire. An important lesson from the pre-digital age is that neither of these constituents—the producers and the preservation entities—necessarily has the mandate or the resources to transfer objects into the library. For certain classes of “traditional” materials, particularly books and journals, third-party service providers have been essential partners to materials acquisition and processing. The same service model might prove most reliable, robust and cost-effective for digital acquisitions and ingest.

As documented by the British Library (BL) in their comprehensive study of life cycle management, preservation investments *begin* at acquisition, with managerial and service interdependencies throughout the information life cycle ultimately dictating the total costs to support any n year life span.² Acquisitions processing (e.g., binding, stamping and labelling) and storage costs are applicable regardless of use; additional downstream investments for conservation or reformatting are inevitable for materials that are in high demand. Although the question remains open of whether traditional formulas to calculate long-term preservation costs will be applicable to digital collections, two conclusions appear to be self-evident: minimizing preservation costs at any given stage, and maximizing the intervals between “interventions” are effective methods to control costs.

Because such preservation interventions will be more frequent to achieve comparable life spans for digital resources, it is imperative to minimize the costs of processing electronic resources at the point of acquisition. Automation will be essential, as will domain expertise for the

intellectual (manual) tasks necessary to select, interpret and prepare materials for deposit to an archive.

The Preservation Market

Commercial ingest tools and services will not be developed if industry does not see a defined market. The consumer market is potentially large in number: virtually everyone with a digital camera or video recorder, for example, will potentially create images she or he wishes to keep for a lifetime. Yet how much will these individuals want to spend on archiving, and how much time would they be willing to invest to organize their images for transfer to an archiving service? For them, convenience will be paramount.

National archives, national libraries, and research libraries constitute a comparatively small market, yet their annual budgets for preservation are relatively large. The high costs associated with conservation, rescue (e.g., digital archaeology) or replacement all serve to motivate libraries to invest in preservation upon acquisition.

The Association of Research Libraries annual *Preservation Statistics* document when and to whom North America's largest libraries disburse preservation funds. In 2001-2002, for example, 116 reporting ARL libraries collectively spent over \$96,500,000 on preservation. Salaries and wages were the largest expenditure (55%), but a significant portion of preservation income (38%) was spent on contract services.

Libraries outsource binding (re-binding of paperback and hardcover books), conservation, microfilming, digitization and "other" services. Given the high volume of acquisitions in these libraries, perhaps it is not surprising to find that binding is the largest contract category, receiving \$25 million (26% of all preservation costs). What is notable, however, is the cost gap between binding and the second-largest category of contract services. Microfilming expenditures (~\$4.7 million) were less than one-fifth of those for binding.³

Statistics from other years mirror this trend: libraries make large aggregate preservation investments, at low per-unit costs, when materials are acquired, to avoid paying higher costs later in the life span to restore usability to items and collections.

The commercial bindery/library relationship might be studied as an example to replicate in creating and marketing digital acquisitions and ingest services. Service providers that provide technical services in document production (e.g., binding books, or encoding e-journals for publishers) are well positioned to process these same materials when they are transferred to libraries and archives for preservation. While the research libraries have the preservation incentives and income, their contract partners might prove to be the best target market for ingest applications.

Overview of Harvard University Library Digital Repository Service

The Harvard University Library (HUL) Digital Repository Service (DRS) is a managed repository for "library-like" digital materials that "support research, have persistent value, and are expected to be on deposit indefinitely."⁴ Harvard owners of "library like" material began depositing objects in 2001.

Over the past three years, approximately two dozen owners—organizational entities such as libraries, archives and museums—have deposited nearly 500,000 objects. At present, nearly all DRS content has been created internally by digitizing library and museum collections, yet DRS policy and systems are well-configured to accept deposits of born digital material from internal and external entities.

Based on a partial cost recovery model, DRS funding is divided into three components with different sources of support:

- DRS infrastructure is considered a common good and is financed through the general "HOLLIS assessment," an annual billing algorithm based on each school's average number of uses of HOLLIS, the library's union catalog of 9 million records to books, journals and other materials. DRS Infrastructure costs include: development staffing, maintenance, operations, and various University Information Systems costs (except those directly related to storage).
- DRS storage costs are recovered by charging fees to the object owner, at the current rate of \$5.00 per GB per year. These storage costs include disk space purchase, maintenance, or rental.
- DRS object transformation costs will also be recovered by charging fees to the object owner. These fees will recover the transformation costs incurred when technological changes require that the DRS transform objects (or their component parts) to a new format to preserve usability. Costs will vary depending on the type of transformation and the number of objects involved.

DRS charges no fees to deposit objects or to use the "DRS Web Admin" tool to perform maintenance on deposited objects and associated metadata.

A deposit to the DRS includes digital objects (ideally in "preferred formats") and their associated administrative and structural metadata. Deposits are submitted via FTP to a drop box on a secure server. In the deposit process, object metadata is carried by a batch load file (called batch.xml), formatted in XML according to the DRS batch DTD. (The batch.xml incorporates mandatory "type" metadata for audio, image, and text formats—with the types to be expanded to accommodate other format classes as needed.) Once validated, objects and metadata are moved from the drop box into the repository.

Deposit Agents

In the DRS lexicon, anyone who deposits objects is referred to as a “depositing agent.” In practice, however, over two-thirds of DRS object owners purchase or delegate deposit services from digitization labs with programming resources. Librarians comfortable with catalogs and other complex databases for descriptive metadata have been daunted by DRS batch.xml in particular, and preservation metadata in general. Clearly, this has been due to lack of tools and accompanying training and support.

Functional Requirements for Ingest

The current cycle of format analysis and applications development is too long at Harvard and other emerging digital archives to keep pace with the needs to deposit at-risk content to safe-harbor repositories. Industry participation in developing robust ingest applications will greatly facilitate the scaling of archiving services.

One caveat is that applications designed to perform ingest functions must be configured to balance automated and manual operations.

Data Marshalling and Selection

In what the *Producer-Archive Interface Methodology Abstract Standard* refers to as the “preliminary phase” of archiving, one must identify the primary information the archive must preserve.⁵ There are cases where multiple versions of content are available—drafts and final versions; raw and processed still images; archival and production audio masters—and the “best” edition(s) to preserve may not be self-evident to the creator, deposit agent (provider of ingest services), or the archive.

As Harvard learned in discussions with e-journal publishers regarding advertisements and other content ancillary to journal articles, “[d]eciding what of all that is seen on e-journal sites today should be archived and maintained will require careful consideration by archives, publishers, and scholars.”⁶ These selection and evaluation tasks cannot be automated, although the decisions could be documented in a *Submission Agreement* or possibly encoded in the administrative metadata that accompanies the object.

Format Analysis

The Library of Congress is developing an analytic model that characterizes the sustainability, quality and functionality of digital formats for archiving.⁷ This framework will likely be incorporated in deposit policies for individual archives, as well as the structure of the documentation for formats registered in the proposed Digital Library Federation Global Digital Format Registry (GDFR).⁸ Key underpinnings of the Library of Congress Model and the GDFR are that format identification, validation and characterization will serve to document integrity, monitor obsolescence and inform the selection of preservation methods least likely to introduce loss.

Format Transformation

Because preservation activities occur at the level of format as well as object, digital archives should be expected to have stated policies distinguishing supported from unsupported (or unacceptable) formats. The implication of such policies is that unconfirming formats will need to be transformed prior to ingest.

Given that format translations introduce risks of losing information or altering an object’s significant properties (underlying meaning), decisions regarding format normalization must be carefully considered—optimally by domain experts such as audio engineers, encoding specialists, digital photographers, etc.

Metadata Production

Metadata production is the heart of pre-ingest activities. Documentation of technical attributes, rights and ownership are essential to preservation. Applications can be designed to parse files for many “self-describing” attributes, but others require human interpretation. How, for example, can one document an object’s significant properties (underlying meaning)? Who decides; by what process? How much of this documentation can be encoded? And without encoding, what value would such documentation provide?

Good ingest applications will serve two purposes: to perform automated extractions, mappings, or even creation of technical metadata to repository-compliant, or standards-based formats; to enable users (i.e., owners and/or deposit agents) to interpret objects—or large classes of objects—then document the artifactual, contextual, evidential or other values that need to be factored in preservation.

Packaging

Single content objects, such as audio performances or digital books, frequently comprise many files with meaningful relationships. How can these be reliably packaged for transfer to and disassembly at the archive? The Metadata Encoding and Transmission Standard (METS) Schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the W3C XML schema language.⁹

Harvard is one of several institutions using METS as the container for packages transferred to repositories.¹⁰ Example applications include: biomedical image “stacks” and accompanying metadata, complex audio preservation packages, and, in the near future, component parts of multi-page objects (e.g., digitized books).

Submission/Deposit

Final steps in the ingest process are object transfer to the archive, via FTP or other means, then reliable and trustworthy deletion of local copies of objects reported to have been deposited successfully to the archive.

Harvard Tools

In keeping with University Library policy to develop infrastructure in response to community priorities, Harvard's Office for Information Systems has overseen the development of applications to facilitate pre-ingest validation, ingest (deposit), and subsequent validation and packaging of data within the archive or disseminated from it. As described below, these applications are currently configured for text, XML, PDF, still images and, digital audio—the major formats currently being deposited to DRS.

JHOVE

To meet repository needs to verify the integrity of deposited objects, JSTOR, the scholarly journal archive, and the Harvard University Library collaborated to develop an extensible framework for format validation. JHOVE (pronounced “jove”), the JSTOR/Harvard Object Validation Environment,¹¹ provides functions to identify, validate, and characterize digital objects.

- Format *identification* is the process of determining the format to which a digital object conforms; in other words, it answers the question: “I have a digital object; what format is it?”
- Format *validation* is the process of determining the level of compliance of a digital object to the specification for its purported format, e.g.: “I have an object purportedly of format F; is it?”

Format validation conformance is determined at two levels: *well-formedness* and *validity*. A digital object is well-formed if it meets the purely syntactic requirements for its format. An object is valid if it is well-formed and it meets additional semantic-level requirements.

For example, a TIFF object is well-formed if it starts with an 8 byte header followed by a sequence of Image File Directories (IFDs), each composed of a 2 byte entry count and a series of 8 byte tagged entries. The object is valid if it meets certain additional semantic-level rules, such as that an RGB file must have at least three sample values per pixel.

- Format *characterization* is the process of determining the format-specific significant properties of an object of a given format, e.g.: “I have an object of format F; what are its salient properties?”

The set of characteristics reported by JHOVE about a digital object is known as the object's representation information, a concept introduced by the Open Archival Information System (OASIS) reference model.¹¹ The standard representation information reported by JHOVE includes: file pathname or URI, last modification date, byte size, format, format version, MIME type, format profiles, and optionally, CRC32, MD5, and SHA-1 checksums.

JHOVE is designed as a layered architecture with an API (with well-defined, public interfaces) invoked by a thin application layer for a stand-alone, command line tool,

applicable for batch and interactive operation. The API can be used on its own to create other compatible tools.

Identification, validation, and characterization actions are frequently necessary during routine operation of digital repositories and for digital preservation activities. These actions are performed by *modules*. The output from JHOVE is controlled by *output handlers*. JHOVE uses an extensible plug-in architecture; it can be configured at the time of its invocation to include whatever specific format modules and output handlers that are desired. The initial release of JHOVE includes modules for arbitrary byte streams, ASCII and UTF-8 encoded text, GIF, JPEG, and TIFF images, PDF, and XML; and text and XML output handlers.

Harvard is currently considering a number of revisions to DRS policy, including stipulating that in order to be accepted to DRS, a format must be listed in the Global Digital Format Registry (or similar persistent registry) and that a JHOVE module for the format is available for pre-ingest validation.

Dmart

Digital audio represents the most complex digital resource that the Harvard University Library has been asked to create, preserve and deliver. DRS preservation audio deposit packages consist of multiple versions of digital audio files—high-resolution archival and production masters, and lower resolution use copies—as well as a wealth of metadata. Audio technical metadata capture the technical properties of the audio files and their processing history. Structural metadata define the relationships among these various components.

To facilitate DRS deposit of audio preservation packages, OIS worked in conjunction with David Ackerman, Audio Preservation Engineer at the Loeb Music Library, to develop the DRS METS Archive Tool (Dmart).¹² Dmart, a desktop application, automates the complicated packaging of audio preservation components by reading a source directory of audio files and produces an audio METS deposit package, *mets.xml*, and a *batch.xml* file for DRS deposit.

The Dmart tool processes an audio source directory and produces a DRS deposit package. Dmart includes a configuration file permitting the deposit agent to input required base metadata (DRS *batch.xml*) documenting ownership, rights and other administrative information pertaining to the object being deposited. The deposit package includes:

- *batch.xml*
- *mets.xml*, with encapsulated AES *audioObject* and *processHistory* files (*audioobject.n.xml*, *digiprov.n.xml*), AES31 Audio Decision List files (*aes31.adl*), and miscellaneous processing files; and pointers to external files:
- Archival master audio, *name.n* or *name.n.wav*
- Production master audio, *name.n* or *name.n.wav*
- Deliverable RealAudio files, *name.n.ra*
- Waveform reduction files, *name.l.r* or *name.gpk*
- SMIL files, *name.smi*

Technical metadata properties about the audio files are extracted from the audioObject.n.xml files and copied into the appropriate fields of the audioMetadata block of the batch.xml file. The deposit package may then be transferred to DRS.

Conclusion

Digital archives will need to keep pace with content producers in order to remove constraints on submission policies that explicitly prohibit formats from being deposited, or, alternatively, from receiving the highest levels of preservation service. Industry-developed applications would be a significant boon to archiving, potentially removing barriers to ingest. Research libraries comprise a key market for these applications, given their perpetual need to acquire and process collections, their preservation budgets, and, in some cases, their operational oversight of digital archives.

Well-designed ingest applications are very likely to prove extensible to other points in the information life cycle: validation, object monitoring and transformations within the archive; packaging and validation at the point of dissemination.

References

1. *Archival Workshop on Ingest, Identification, and Certification Standards* (1999), which identified ingest as one of the "most urgent areas" in digital archiving requiring work.
2. Helen Shenton, *Life Cycle Collection Management*, LIBER Quarterly, 13, no. 3/4 (2003).
3. Association of Research Libraries, *ARL Preservation Statistics 2001-02* (2003).
4. Harvard University Library, *DRS Policy Guide* (2001).
5. Consultative Committee for Space Data Systems (CCSDS), *Producer-Archive Interface Methodology Abstract Standard, CCSDS-651.0-R-1, Red Book* (2002).
6. Dale Flecker, *Preserving Scholarly E-Journals*, D-Lib Magazine, Vo. 7, No. 9 (2001).
7. Caroline R. Arms and Carl Fleischhauer, *Digital Formats for Library of Congress Collections: Factors to Consider When Choosing Digital Formats*, November 7, 2003 DRAFT.
8. Stephen L. Abrams and David Seaman, *Towards a Global Digital Format Registry*, World Library and Information Congress: 69th IFLA General Conference and Council (2003).
9. Metadata Encoding & Transmission Standard, *METS Implementation Registry* (2003).
10. Harvard University Library, *JHOVE Format-Specific Digital Object Validation* (2004).
11. ISO/IEC 14721:2002, *Space data and information transfer systems -- Open archival information system -- Reference Model* (2002).
12. Harvard University Library, *DRS METS Archive Tool (Dmart) for Audio Deposit* (2003).

Biographies

Stephen Chapman is Preservation Librarian for Digital Initiatives in the Weissman Preservation Center, Harvard University Library. He advises curators and other members of the Harvard community about approaches to collections digitization, and facilitates discussions and investigations among technology developers and digitization practitioners to develop efficient systems and workflows for creating and archiving digital library materials. Mr. Chapman co-authored the draft data dictionary for the NISO Z39.87-2002 Standard, Technical Metadata for Digital Still Images.

Stephen Abrams is the Digital Library Program Manager at the Harvard University Library, providing technical leadership for strategic planning, design, and coordination of the Library's digital systems, projects, and assets. He is currently engaged in research and implementation of effective methods for archival preservation of digital objects. Mr. Abrams is the ISO project leader and document editor for ISO/TC 171/SC 2/WG 5, the joint working group developing the PDF/A standard. He is a member of ACM, ALA, SIS&T, and the IEEE Computer Society.