The Archive Ingest and Handling Test: Implications of Diverse Content and Diverse Repository Practices

Martha Anderson
Library of Congress
101 Independence Ave., SE
Washington, DC 20540
(202) 707-2598
mande@loc.gov

Bill LeFurgy
Library of Congress
101 Independence Ave., SE
Washington, DC 20540
(202) 707-8618
wlef@loc.gov

ABSTRACT
In this paper, we describe a recent Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) project to document and test exchange of content among digital repositories. Known as the Archive Ingest and Handling Test (AIHT), the year-long project involved the transfer, ingest, and export of a test data set amongst the Library and four partner universities (Harvard, Stanford, Old Dominion, and Johns Hopkins), using a variety of digital repository technologies and approaches.

Categories and Subject Descriptors
H.3.2—Information Storage; H.3.4 Systems and Software – *Performance evaluation (efficiency and effectiveness)*, *Distributed systems*; H.3.7 Digital Libraries – S*tandards*, *Systems issue*

General Terms
Documentation, Design, Experimentation, Standardization, Theory

Keywords
Archive Ingest and Handling Test, AIHT, digital preservation, repositories

1. INTRODUCTION
What are the practical issues associated with placing rich content into different digital repositories? To what degree are repositories similar in how they go about managing content? Are processes standardized enough to enable broadly based metrics? How does a repository measure success?

Despite the attention directed in recent years toward the use of repositories for digital preservation, answers to such questions remain elusive. But they must be addressed before we can develop collective confidence in undertaking long-term stewardship of digital collections. Getting answers requires tests of current systems: We need to know what actually is involved in placing representative content into operational repositories. The results will help point to improved methods, both locally and globally.

This was the intent of a recent Library of Congress, National Digital Information Infrastructure and Preservation Program (NDIIPP) [1] project to document and test moving content among digital repositories. Known as the Archive Ingest and Handling Test (AIHT), the year-long project involved the Library working with four universities with a variety of digital repository technologies and approaches: Harvard, Stanford, Old Dominion, and Johns Hopkins. As detailed by Clay Shirky in the December 2005 issue of D-Lib Magazine [2], AIHT required each repository to transfer, ingest, and export a test data set. The test set was the George Mason University (GMU) 9/11 Archive [3], consisting of about 12 gigabytes with 57,000 digital files collected in connection to the September 11, 2001, terrorist attacks. The file formats were mixed, with most associated with email and the Web. Formats fell into three basic categories: plain text, HTML text markup, and audiovisual, primarily images. There was no overall metadata scheme; in fact, most individual files had little or no descriptive metadata. Much of the content had been submitted singly by individuals. The short duration of the AIHT project and the gaps in metadata led the participants to concentrate on bit preservation with some secondary attention to object preservation where local repository tools and methods allowed.

The test data set presents a real-world dilemma. The content is of clear value for documenting a significant event and the impact of that event on society. It also has value as evidence for how technology was used to convey and capture information of widespread interest. As such, the collection warrants preservation. But the collection was not originally assembled with preservation in mind. The data is heterogeneous and imperfect, with varied formats, limited descriptive details, and no way to get better information from content creators. This mix of high value and deep technical challenge reflects contemporary reality in terms of how many digital items are created and used. As they take root and grow, many digital repositories will inevitably confront collections with these characteristics.

2. PRACTICAL TESTS
The AIHT was designed to obtain practical information about the steps involved for a repository to manage a mixed collection of digital content that was assembled independent of existing submission guidelines. A true "stress test," the AIHT intended to illuminate the degree to which current repository practices and technology can accommodate a rich, varied, and non-standardized body of digital content.

Joint Conference on Digital Libraries (JCDL) 2006, June 15, 2006, Chapel Hill, NC, USA.

Test results were sobering. There were a myriad of low-level problems that appeared immediately: file names changed when moved from the original operating system to intermediate systems; file counts varied; metadata—when present—was uneven; and file formats on occasion proved difficult to confirm. In short, the underlying fragility of the content itself and the tools and systems used to manage it were both exposed. This points to a compelling need for rethinking assumptions about repositories.

As Shirky notes in connection to the AIHT, "Requirements Aren't." The repositories participating in the test had detailed policies and requirements for file formats, metadata, donor information, and so forth that had to be met before content could be ingested and preserved. The AIHT rendered all of these moot. Working with the content necessitated bending or ignoring many preconditions that were put in place to simplify ingest. Participants concluded that a collection like the 9/11 Archive compels what is in effect triage: Using a customized methodology to generate essential metadata and "clean up" files as needed. This process also required weighing the risk of data loss against the effort needed to mitigate loss.

3. IMPORTANCE OF HETEROGENEITY

The test also revealed the significance of what can be called repository heterogeneity. Each repository depended upon its own framework of business rules to operate. These rules stemmed from, and were enmeshed within, a larger set of institutional rules, regulations, mandates, and histories that were entirely unique. This influenced everything from conceptualizing the test set as a whole to checking individual files. There were, for example, different ideas about the 9/11 Archive data model. Some institutions viewed it as one entity, while others viewed it as a conglomeration of many small entities. Similarly, each repository chose to frame preservation tasks in its own fashion, with relatively little reliance on or trust in what others had done with the content beforehand. The participants chose to place their trust in their own local tools and processes, which they understood and implemented within the context of their own organization. Even in cases where a single tool was used by all (such as the JSTOR-Harvard Object Verification Engine, or JHOVE) [4], each institution's individual culture drove use of the tool (and interpretation of its results) in different ways.

The importance of heterogeneity is demonstrated by the processes each repository used to export the test data set for another round of ingest by a peer. Staff from the repositories knew each other well and had been working together on the AIHT project for almost a year. Certain tools, such as JHOVE, were used by all. And yet when each repository exported its own validated version of the test data set, neatly packaged in accordance with community standards, complete with metadata derived from analysis of individual files, that package was regarded much the same way as the original raw data. Metadata generated from one repository's use of JHOVE was essentially disregarded, as each felt obligated to rerun JHOVE to obtain trusted results.

Heterogeneity offers value while also imposing a cost. It is useful to avoid relying on a monoculture in dealing with complex situations, as diversity provides for varied paths forward while building in a defense against mistakes that only become apparent later on. These issues have particular resonance at the current stage of development for digital preservation. Our knowledge and practices are still limited and it is risky to think that we now know the best way to do things or to have confidence in a direct course to specific outcomes or solutions. Reliance on a uniform approach may result in the loss of important information, especially if that approach relies on overly strict requirements for the data or its metadata. There is also a danger that digital information will be lost if institutions wait for development of optimal preservation technology that has some perceived broad applicability. But heterogeneity can be seen as imposing less efficiency and more cost to the overall preservation effort. Conceptually, at least, a more uniform set of processes and standards might cut development time and lead to more easily deployed systems.

A question that arises from this relates to the future of repository homogeneity: Is it going to decrease with shared knowledge and better systems, or is it going to increase with more institutions engaging in preservation in accordance with unique business rules? It may be too soon to predict. As noted, we are still in the early days of digital preservation. What seems clear at this point, however, is that an institution that operates a digital repository does so to meet its own particular needs and in its own trusted fashion. There is a limited institutional expectation that other repositories are "doing it right," even if they are using the same tools and have a similar conceptualization of the issue.

Repository heterogeneity may, however, coexist in harmony with a high level of interoperability and shared infrastructure. There is, for example, a need for much more robust methods to distribute and keep copies of content and to support ingest (or even "re-ingest"), as well as redundant storage [5]. The widespread use of JHOVE and protocols such as the Metadata Encoding and Transport Standard (METS) [6] indicates a clear market for shared tools and standards—even if they are used in an embedded local context. Perhaps our notion of interoperability will evolve to include a concept of something like a "repository agnostic" infrastructure layer that has sub-sections that can be used—or not—in accordance with local needs. Another metaphor might be that of a city: There is a generic infrastructure made up of roads, electrical power, water, and fire and police protection. This permits individuals and businesses to use the infrastructure to more easily build specific structures for unique needs.

Regardless of the direction that repository heterogeneity goes, it will be necessary to develop some means for determining preservation "success." This will have to be done with an eye to loss of data and of functionality. Given the fragility of data, and of our systems, some information will inevitably be lost, changed, or corrupted despite our best efforts into the foreseeable future. We need to determine what levels of loss are acceptable, be it 1 file in 10, 1000, or 1,000,000. There also needs to be more forthright acknowledgement and discussion of data loss in connection with preservation efforts.

Transparency in terms of repository policies and selective outcomes (such as ability to access information at some specified level of service) is also important. Terms of transparency may turn out to be best negotiated and enforced in a consortial context, which in turn could depend on varied legal, contractual, or business models. Given all these variables and uncertainties, it may be too early to rely on a uniform process to broadly assess or measure repositories in any absolute sense.

4. PRACTICAL VALUE

At the end of the AIHT project, participants acknowledged the value in the practical exercise for their own institutions as well as for lessons learned that may benefit the preservation community. They also came to understand concretely the challenges of working with the current transfer protocols such as FTP and operating system tools for the purposes of preservation at large scale. As a group, the project teams articulated and tested processes for gaining control of and managing a large and diverse data set. One of the most promising outcomes has been that the participants achieved a common understanding that more work is needed to clarify and document the most useful standard interfaces for transfer and ingest of content collections between diverse repository systems. Interoperability will depend on such interfaces to bridge the unique needs, policies, and processes that institutions depend on to manage digital preservation repositories.

5. ACKNOWLEDGMENTS

Our thanks to Clay Shirky for his help, and also to the AIHT project leaders and their teams: Stephen Abrams, Harvard University; Sayeed Choudhury and Tim DiLauro, Johns Hopkins University; Keith Johnson, Stanford University; and Michael Nelson, Old Dominion University. Thanks also for the editorial assistance of William "Butch" Lazorchak.

6. REFERENCES

[1] *Digital Preservation (Library of Congress).* Available at http://www.digitalpreservation.gov. Accessed June 2, 2006.

[2] Shirky, C. AIHT: Conceptual Issues from Practical Tests. *D-Lib Magazine, Vol 11. No. 12, December 2005.* Available at http://www.dlib.org/dlib/december05/shirky/12shirky.html. Accessed June 2, 2006.

[3] *The September 11 Digital Archive.* Available at http://www.911digitalarchive.org/. Accessed June 2, 2006.

[4] *JHOVE - JSTOR/Harvard Object Validation Environment.* Available at http://hul.harvard.edu/jhove/. Accessed June 2, 2006.

[5] Bekaert, J. and Van de Sompel, H. A Standards-based Solution for the Accurate Transfer of Digital Assets. *D-Lib Magazine, Vol. 11, No. 6, June 2005.* Available at http://www.dlib.org/dlib/june05/bekaert/06bekaert.html. Accessed June 2, 2006.

[6] *Metadata Encoding and Transmission Standard (METS) Official Web Site.* Available at http://www.loc.gov/standards/mets/. Accessed June 2, 2006.