Library of Congress
Archive Ingest and Handling Test (AIHT)
Final Report

Clay Shirky
June 2005

DigitalPRESERVATION
National Digital Information Infrastructure & Preservation Program:
A Collaborative Initiative of the Library of Congress

Table of Contents

Section 1: Introduction

This report documents the development, administration and conclusions from the Archive Ingest and Handling Test (AIHT), a multiparty test of various digital preservation regimes, conducted under the auspices of the National Digital Information Infrastructure and Preservation Program (NDIIPP). It describes the genesis of NDIIPP and of the AIHT under the direction of the Library of Congress; details the phases of the AIHT, documents lessons learned during the test and suggests possible fruitful areas of future work.

This section of the report covers the background of the NDIIPP project in general and the design of the AIHT.

Section 2 is an executive summary of the three main phases of the test -- ingest and markup of a digital archive; export and sharing of that same digital archive from the tested preservation regimes; and conversion of digital objects from one format to another. It also includes observations made in the start-up phase, prior to the official start of the AIHT; overall descriptions of the lessons learned; and suggests future areas of work on some of the issues uncovered in the test.

Sections 3-6 are the principal contents of the report, being the final reports of the participating institutions: Harvard, testing the Harvard Digital Repository; Johns Hopkins, testing both DSpace and Fedora; Old Dominion University, designing "self-archiving objects" based on Fedora; and Stanford, testing the Stanford Digital Repository. These reports contain the description and insights from the four institutions conducting the test and include observations specific to those institutions and preservation regimes.

There is a considerable amount of detail in the participant reports, but there are three significant messages that have come through from this work and which we believe point to valuable areas of future investigation: validation of shared effort, a need for regular, comparative testing, and work to reduce the cost of sharing the preservation burden.

Validation of Shared Effort

The AIHT was created with the idea that the problem of digital preservation was really a complex of problems, strongly affected by the content being preserved, the audience for whom it was being preserved, and the institution doing the preserving. As a corollary, the AIHT (and indeed all the work of NDIIPP) is being conducted with the idea that large-scale digital preservation will require continuing participation of many entities.

The AIHT was designed in part to test this hypothesis of shared value. By designing a test in which the participants exported files to one another, we were able to simulate, albeit on a small scale, some of the shared effort we believe will be required for varying types of institutions to operate in a loose federation.

In particular, Phase 2 of the AIHT tested the ability of various preserving institutions to generate value for one another by sharing full, exported versions of archives, even when those institutions generate different sorts of metadata and use different tools and processes internally. The results of that sharing were generally heartening, with some convergence on metadata standards (METS was independently adopted by three of the four participants), and, perhaps more important, an increase in the understanding and trust among the participants.

We only began the effort required to share whole archives in the second half of the AIHT after all of the participants had had a chance to handle the archive on their own and had got to know one another through meetings, phone calls and e-mail. It is clear that the subsequent requirement to export the entire archive and import another institution's version was greatly improved by these conversations, even with a relatively limited amount of contact. This suggests that forums for comparison and sharing of technical problems and approaches, in any multi-participant environment, will help catalyze both informal and formal networks for shared handling of preservation issues.

Need for Regular, Comparative Testing

The second big lesson from the AIHT was a pressing need for continual comparative testing of preservation tools and technologies. The four participants tested five preservation systems -- Harvard's Digital Repository, Stanford's Digital Repository, DSpace and the Fedora framework twice, once by Johns Hopkins and once in Old Dominion's Buckets, a term referring to self-archiving objects that travel in a bundle with their own metadata.

In all cases, there were significant early issues that were conceptual, which we expected, and technical, which we did not. File names were silently altered during transfer. Archives optimized for one-at-a-time ingest were hard to adapt to bulk ingest. Archives whose performance was well tuned for holding tens of thousands of objects degraded precipitously at hundreds of thousands of objects. And so on.

In fact, every phase of the test exposed significant low-level issues. Institution-scale digital preservation tools are new and have typically been tested in only a few environmental settings, generally in the environment of their creation. No matter how rigorous these tests may be, such a situation creates a risk of lack of portability, at the least. (The Johns Hopkins work was notable in that the institution is one of the first to work on digital preservation in a comparative context, testing existing architectures rather than designing a new one.)

Ten years after the Web turned every institution into an accidental publisher, the simple difficulties of long-term storage are turning them into accidental archivists as well. For digital preservation to flourish, those institutions must be able to implement preservation tools without having to create them from scratch. Though it is still far too early to imagine a guide to selecting preservation tools based on institutional capability and need, we need to keep such a goal in mind.

Continual testing of tools for managing digital content, and publication of the results, will be critical to driving adoption of such tools. This will bring about several obvious

benefits: Continual testing is an essential precursor to continual improvement. A steady stream of bug reports and feature requests will be valuable input to the creators and modifiers of such tools. A well-known public test site or network of such sites will create an environment where the creators of preservation tools can assess one another's work, including examining the interfaces such systems will need to support in any larger federation of preserving institutions. Finally, the ability to study the platforms being tested will begin to give potential users of such systems a view into what is available and appropriate in various contexts.

Reduced Cost of Sharing the Preservation Burden

Finally, all models of shared participation among preserving institutions require that such sharing not bankrupt the participants. Although the cost per unit of storage is still falling dramatically, that change alone is not enough to make digital preservation cost-effective. The steady decline in storage prices cuts both ways, making the creation of enormous new archives possible, and much of the cost in preserving digital material is in the organizational and institutional imperatives of preservation, not the technological ones of storage.

Institutions will only make the choice to share the burdens of preservation when they can do so without having to bear crippling costs, either financial or human. Given how much of the cost is in the initial design and implementation of a working architecture, anything that helps institutions adopt and effect digital preservation regimes will accelerate the spread of such systems. In addition, shared investment in handling digital data, in a manner analogous to union catalogs or shared collection building, will benefit from any efforts to simplify the sharing of marked-up data between institutions. The general principle is that reduced cost is an essential prerequisite to the spread of effective digital preservation; that shared effort in definition of such systems is a good way to lower those costs.

Background of NDIIPP

The preservation of digital content has become a major challenge for society. In 1998 the Library of Congress began to develop a digital strategy with a group of senior managers who were charged with assessing the roles and responsibilities of the Library in the digital environment. This oversight group was headed by the Associate Librarian for Strategic Initiatives, the Associate Librarian for Library Services and the Register of Copyrights. This group has held several planning meetings to assess the current state of digital archiving and preservation. The Library has also assembled a National Digital Strategy Advisory Board to guide the Library and its partners as they work to develop a strategy and plan, subject to approval by Congress.

At the same time, Librarian of Congress James H. Billington commissioned the National Research Council Computer Science and Telecommunications Board of the National Academy of Sciences (NAS) to evaluate the Library's readiness to meet the challenges of the rapidly evolving digital world. The NAS report, *LC 21: A Digital Strategy for the Library of Congress*, recommended that the Library, working with other federal and nonfederal institutions, take the lead in a national, cooperative effort to archive and preserve digital information.

Congress appropriated roughly $100 million for a national digital-strategy effort, to be led by the Library of Congress. The Library was chosen not only because of its mission to "sustain and preserve a universal collection of knowledge and creativity for future generations," but also because of its role as one of the leading providers of high-quality content on the Internet.

The effort, named the National Digital Information Infrastructure and Preservation Program (NDIIPP), requires the Library to develop and administer programs that advance the cause of digital preservation through a variety of initiatives and partnerships. The work of NDIIPP is not merely to ready the Library of Congress for the 21st century work of preserving our digital heritage, but to create partnerships with commercial, nonprofit and other government agencies at the federal, state and local levels in such a way as to

improve the state of digital preservation in the country generally. In addition to the Archive Ingest and Handling Test, the subject of this report, NDIIPP has several efforts under way, which are documented at http://www.digitalpreservation.gov.

Discovery Phase

NDIIPP began with a discovery phase. The Library of Congress engaged the Global Business Network (GBN) to convene groups of librarians, archivists, technologists and other stakeholders to discuss the problems in digital preservation and how to approach them. Several meetings were convened both before and after Congress approved the creation of NDIIPP. The conversations at those meetings were wide-ranging and valuable across a number of fronts.

Three things in particular became clear from those meetings. First, many groups were working on the digital preservation problem from their own perspectives, and those groups had particular points of view honed from decades of serving their own users. Those points of view were both valuable and incompatible.

This fact kept us from fantasizing about a one-size-fits-all approach. Instead, we concluded that any proposed digital preservation strategy must describe the minimal set of required functions in such a way that existing systems can be mapped onto the architecture and vice-versa. A survey of the existing literature makes it clear that though there is a common subset of functions required for preserving digital materials, the arrangement and even the names of those functions differs from system to system. Our discussions with practitioners also made it clear that those differences are likely to increase in the near future, as the number of institutions that find they need to preserve their born-digital materials increases.

The second thing that became clear in those meetings was that if the various groups working on digital preservation did not cooperate, there would be an enormous amount of duplicated effort. These two points taken together -- the significant autonomy of

preserving institutions and the value in cooperation -- suggested a model of federation, where preservation becomes a function not just of individual institutions, but of a cooperating network.

The third thing that became clear was that most institutions working on digital preservation were thinking of themselves as endpoints in the passage of digital material from donors to preservation environments. This was difficult to reconcile with a goal of shared effort, since some sort of sharing between preserving institutions would be key to ensuring that digital material was distributed in multiple, geographically separated copies, stored in dissimilar preservation regimes, all of which would serve to mitigate the risk that disappearance of any one preservation environment would be fatal.

With these three realizations, and given the tension between them, it became apparent that the idea of shared effort needed some sort of controlled test, both to observe the operation of different digital preservation regimes in different environments and to examine what would be required to move a digital archive from one such environment to another.

As these meetings were progressing, the Library of Congress and George Mason University (GMU) were separately discussing a donation of GMU's 9/11 Digital Archive (www.911digitalarchive.org/) of content related to the terrorist attacks of Sept. 11, 2001, and their aftermath. Because this put the Library in the position of managing the transfer of an entire digital archive, the Library decided to negotiate with GMU for the right to use identical copies of that archive as a test archive, given its advantages as a real-world heterogeneous collection of data.

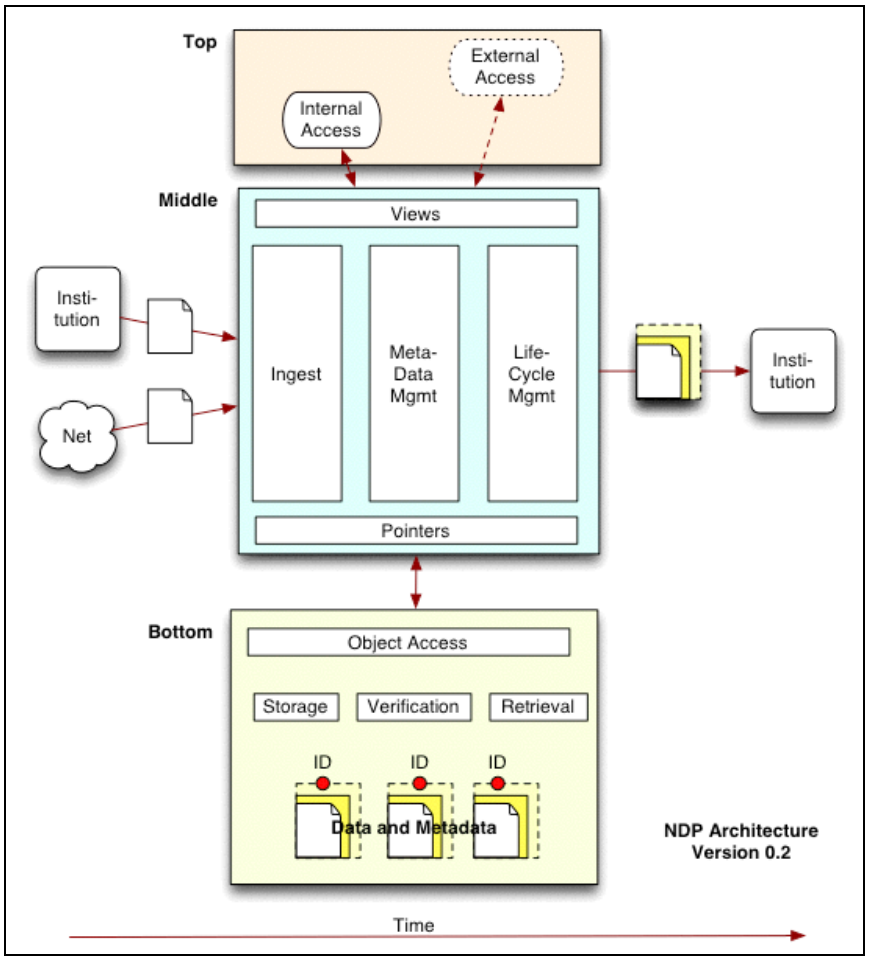
Goals of the AIHT

The AIHT was designed to be the first practical test of interfaces specified in the architectural model for NDIIPP. That architectural detailed in documents available at www.digitalpreservation.gov/technical/ is best thought of as a set of interfaces between

different organizations, whether they are separate institutions, or separate departments or other service providers within a single institution.

The idea behind the NDIIPP architecture is that no one technical design can specify all possible uses, but that there are a handful of common elements, generally involving the transfer of both data and responsibility or control between parties, that can be specified. The common elements are a set of layers -- Top, where access to material is provided; Bottom, where long-term storage of bits takes place, and Middle, which handles the complex of functions related to ingest and management of digital content. There are interfaces between the layers, and there are also interfaces for ingest and export of digital material. The basic interfaces in the architecture are shown in Figure 1 below:

Figure 1. NDP Architecture Version 0.2



The AIHT was meant to test the ingest and export interfaces, which is to say the passage of material from the outside world into a preserving institution, and from there to be exported for re-ingest elsewhere. The architecture is also designed to be modular as to responsibility once the material is ingested; for instance, it is possible for a third party to offer storage services to a preserving institution that still maintains functions in house of ingest and life-cycle management. The AIHT was not designed to test this modularity; those interfaces will require a different test environment.

The functions tested were those of the Middle layer. At the left, digital material passes into a preserving institution, whether donated by a person or institution or automatically accessed as with a Web crawl. In the center are the functions of the preserving institution or institutions. On the right is the export of material from the preserving institution outward. Note that this export can be in frequent small batches or in periodic snapshot exports of an entire collection or archive. Note also that the architecture assumes that data is exported in a format that packages the object with some additional metadata.

The Middle layer contains five large functional categories. The five categories in the Middle layer encapsulate a wide range of functions – any working system will have to break out those functions in more detail. However, a survey of the literature and practice suggests that the next level of detail is where existing systems begin to diverge:

1. **Ingest** – the functions required for the transfer of responsibility for the preservation of digital data to a particular organization, including both acceptance of digital materials and the creation of any contractual or administrative agreements.
2. **Pointer Management** – the creation or registration of pointers for the digital objects being preserved. Pointers point to digital objects stored in the Lower layer.
3. **Metadata Management** – the creation and management of metadata for the digital objects being preserved. (Note that at least some metadata will be stored with the object itself.) At a minimum, this metadata will include or point to as much detail as possible on making the object available for interpretive use – file

format, conditions of creation, playback software, etc. Note that the metadata can be stored by other institutions, including third-party service providers, as well as by the hosting institution.

Note also that additional metadata will be developed over time, in forms ranging from additional management or scholarly annotation to administrative notes related to the management of the object.

4. **Life Cycle Management** – the set of operations required to keep digital data fit for use over the passage of time, including transferring copies of the original objects in bit-identical format onto new storage media; migration of objects to new formats; documentation of emulation strategies for playing back older data on newer software; and exporting objects, which entails the possible transfer of metadata and of preservation responsibility, if it is contractually agreed, to other preserving institutions.

5. **Views** – The Views function essentially plays a gatekeeper role for the provision of access to the objects, filtered through whatever policies or restrictions are placed on their use -- internal only or available to other institutions, any particular file transformations that are allowed or disallowed, etc.

The AIHT tested the first four functions -- Ingest, Pointer Management, Metadata Management, and Life Cycle Management. Because the material being handled was copyrighted, and because it was being subjected to experimental handling, all the participants agreed not to make it available to end users, thus leaving the Views function to cover only ad hoc access by the staffs of the participating institutions.

With the GMU archive as our target test product, and the functions of the Middle layer as the focus of the test, we drafted a Call for Proposals (CFP), to find partners who would work with us on the AIHT.

The Call for Proposals described the AIHT this way:

> The essence of preservation is institutional commitment. Because the number of individuals and organizations who produce digital material is far larger and growing much faster than the number of institutions committed to preserving such material, any practical preservation strategy will require mechanisms for continuous transfer of content from the wider world into the hands of preserving institutions.
>
> The Archive Ingest and Handling Test (AIHT) is designed to test the feasibility of transferring digital archives *in toto* from one institution to another. The purpose is to test the stresses involved in the wholesale transfer, ingestion, management, and export of a relatively modest digital archive, whose content and form are both static. The test is designed to assess the process of digital ingest, to document useful practices, to discover which parts of the handling of digital material can be automated and to identify areas that require further research or development.

The project was designed in three phases. The first phase of the test was primarily concerned with the operation of individual preservation regimes and was designed to test the ingest of a small, relatively heterogeneous real-world archive. After ingest, participants were asked to identify and perform subsequent operations on the digital objects required to sustain them in a form that kept them fit for use, including alternate formats for display in anticipation of format obsolescence.

The second phase of the AIHT involved cooperation between institutions, examining the issues and protocol requirements for the exchange of whole archives or collections, with accompanying metadata of whatever richness, between institutions operating different preservation regimes. The participants in the AIHT were able to develop strategies for this handoff, for things like delivery method and format, based upon lessons learned in the first phase.

The test archive of material related to the events of Sept. 11, 2001 was donated to the Library of Congress by George Mason University. It was chosen for the AIHT because it was a real-world case. In addition, the archive offered two advantages of scale -- it was large enough with approximately 57,000 files in roughly 20 file types to present non-trivial metadata management issues but small enough at 12 gigabytes to present few difficulties in storage or transfer.

The project did not test issues of general public search or access, rights clearance or management, long-term viability of various storage media, security, evidentiary provenance or terabyte-scale ingest.

Harvard University Library, The Johns Hopkins University Sheridan Library, the Computer Science Department at Old Dominion University, and Stanford University Library and Information Resources responded to the call and were selected for participation. With those participants, the GMU archive and goals as specified in the call, the test began in January 2004 and concluded in May 2005.

Section 2 is an overview of issues and observations that arose during the test that were not limited to the individual participants. Sections 3-6 are the reports of the individual participants. A technical description of the GMU 9/11 Digital Archive is included in the Appendix

Section 2: Overview

This section is in several parts, with each part corresponding to one period of the AIHT. It is not intended as a complete accounting of all the work of each of the participants, which is covered in Sections 3-6, but rather as a summary of critical observations, lessons or issues uncovered in each phase of the AIHT.

The sections are:

- **Startup** covers the earliest phase of the program, the transfer of the digital archive from GMU to the Library of Congress.

- **Phase 1: Ingest** covers the initial ingest and markup of the archive.
- **Phase 2: Export** covers the export of a marked-up version of the archive, and re-import by other participants.
- **Phase 3: Conversion** covers the local conversion of digital objects in the archive from one format to another.
- **Conclusion** covers observations that arose during the test but were not specific to any one phase.

It is our hope that the results of the AIHT will continue to inform work on the design and testing of tools and techniques for digital preservation within the Library of Congress, with the participants in the AIHT itself and in the larger preservation community.

Startup

We had imagined that the startup phase, in which the archive was transferred from GMU to the Library, would be extremely easy. In fact, we discovered that even seemingly simple events such as the transfer of an archive are fraught with low-level problems, problems that are in the main related to differing institutional cultures and expectations. Because we had expected the handoff to be trouble-free, the lessons learned during this phase were particularly surprising:

The Simplest Technology Choices Are Often the Most Important

During the initial transfer of the archive from GMU to the Library, we chose to ship the archive on an NTFS file system (Windows), even though the original archive was kept on an ext3 formatted file system (Unix/Linux). This turned out to be the single biggest mistake in the entire transfer process, as the file names as preserved on ext3 were altered automatically and without warning by the NTFS file system, which has different expectations and restrictions on which sort of characters are allowed in file names.

For example, the archive, as expressed on the NTFS file system, included a number of file names that looked like:

ikonfriend.cgiforum1topic3
help.cgihelponRegistering
hpxs5002en.gif1000241013

These file names were preserved from the original URLs of the harvested sites. It is obvious that somewhere the relevant punctuation marks for query strings (?, =, and &) have been deleted.

This had two negative effects -- first, it obscured the actual file extension (e.g., .cgi), by eliding the query string with the file name. Even worse, a file name like help.cgihelponRegistering could have been either of:

help.cgi?helponRegistering
help.cgi?help=onRegistering

Even when the query string is composed of simple words, there is no way to return it to its pre-edited format with complete certainty.

As a result, we decided, as a basic principle, to postpone all conversion issues, even accidental ones, until the information to be preserved was in the hands of the receiving institution. Where possible, taking physical possession of the original disks for subsequent copying would introduce the least distortion outside the hands of the receiving institution. Even operations as seemingly innocuous as burning an archive to a CD before transfer could introduce subtle transformations, and while such operations may be inevitable, staff should be prepared to look for unexpected alterations, especially as the file name or other unique handle for incoming data is one of the key pieces of metadata at any original ingest.

Identifiers Aren't

There were several sorts of identifiers mixed together in the system -- URLs, file names, file extensions, query strings and digests. The URLs were the unique identifiers as taken from sites spidered or donated from the Web. File and directory names were either generated from these URLs, recreating some of the directory structure of the spidered or donated sites, or else were generated upon being added to the GMU collection. File extensions came in several types, including labels of content type (e.g., .txt, .doc); instructions to the original Web server (e.g., .cgi, .asp); version numbers for older, undisplayed content (e.g., .jpg.2), and user-generated notation (e.g., .data.) Query strings are content appended to a URL after a "?" symbol and act as additional instructions to a Web server. But in this instance they were also preserved as part of the file names during transfer, and digests were MD5 checksums of the content.

None of these identifiers worked perfectly, even for the limited service they had to perform during the transfer. URLs were not easily convertible to file names, because of file system limitations (see below.) File extensions were both variable (.jpg, .jpeg, .JPEG) and often pointed to context that cannot be recreated. It is impossible to know the structure of the CGI program if all you have is the resulting output. In some cases the labels were simply incorrect (e.g.GIF files labeled .jpg).

Query strings were similarly altered by the receiving file systems and also suffer from the loss of context in the absence of the interpreting program, and digests, which are guaranteed to point uniquely to collections of bits, only referred to atomic uniqueness, not contextual uniqueness. If 10 separate donors submitted the same text or photo as relevant, the social context of multiple submissions would be destroyed by treating all data with identical MD5 hashes as identical from the point of view of the potential user.

Adding a contextual complication, the original report of file extensions used a naïve parser that regarded any string of letters after the first dot as a file extension, so that

photo.downtown.jpg would have a reported file extension of "downtown." Though this was not an issue with the file names themselves, it demonstrated the potential loss that can come from the interaction of one set of assumptions (file names are all in "name.ext" format) with technological reality (file names can have zero, one or many dots in them.)

As a result, the identifiers that come from donors or other external sources must be examined carefully and cannot simply be taken at face value. Furthermore, during ingest, there need to be ways of independently evaluating the information that may be contained in the identifiers, such as determining whether files labeled .jpg are really in JPEG format.

File Names Must Be Preserved Separately from File Systems

Following the above, because file names are subject to surprising errors in transfer, they cannot be regarded as transparently transmittable between systems. The unpleasant fact is that file systems are not universally forgiving in allowing the transmission of file names, and though file systems ca. 2005 are considerably more interoperable in that regard, the rise of long file names, file names in Unicode and the like means the threat of silent modification of the original file name during transfer will not go away for the foreseeable future. They are, however, a critical piece of metadata, often containing both descriptive and technical components (e.g. eyewitness_account.pdf.)

We can see two obvious responses to the dilemma created when file systems alter file names:

First, all file names should be recorded as metadata someplace besides the file system, ideally wherever the rest of the metadata concerning the particular file is being held. This requires the creation of a second sort of pointer to the file itself, however it is being held, with the file name stored as metadata referencing the digital

object it was previously a handle for. This should happen even if the file names are successfully preserved in a file system as well.

Second, the current file name in any given file system may need to be canonicalized in some way. URL-encoding a file name, which reduces the number of characters that can be used to an almost universally supported subset is one sort of transformation, and one that may be especially useful when archiving Web content. It lengthens file names, however, adding three characters (e.g., spaces are replaced with the three character set %20), meaning that if the designated file system has limits on both character types and length, URL-encoding may be a bad choice.

An alternative possibility is to hash the original file name, and use the result as an independent identifier to the file. (This should be done separately from hashing the file's contents, with or without file name included.) The MD5 hashing algorithm, for example, will return a 32-byte signature using only hexadecimal (0-9 plus a-f) characters. This output is portable across all "modern" (post-DOS) file systems and allows at least positive identification of a match with the original file name (provided, of course, that that has been preserved as well.)

The basic principle is that, at the moment of ingest, it is likely that the file name of a digital object is likely to function both as metadata and as a UID, and that these two functions should be separated and treated separately for all subsequent transfers or transformations of the data.

Context Is Easy to Destroy in Transfer, Especially the First Transfer

In addition to destroying the file names of server-side scripts, the harvesting process of Web sites does nothing to preserve the actual scripts on the server that provide the content. As a result, any complex or context-specific behavior implemented on the server was not captured in the GMU archive. This is a generic problem with Web crawls, and in fact with dynamic content generally, but since the original donation

serves as the first link in a chain of handling the content, any lost context in the first handoff is particularly problematic.

This context-loss problem also happened at the moment GMU ingested some of the content. For instance, although GMU collected self-reported rights for multimedia content, especially photographs, the default position of the form presented to users was checked "yes," meaning there was no way to disambiguate between users who answered the question in the affirmative and users who did not answer the question. As a result, the metadata surrounding rights, though it was collected and transmitted, turned out not to be actionable in determining the rights of the submitted photos, because the default behavior of the system destroyed context.

This makes the initial handoff the riskiest, in many ways, since the donating individual or group is unlikely to have the resources to produce a complete accounting of the context of an even moderately large number of files. (The special case of file-by-file ingest is ideal in many respects, as you can query the donor more thoroughly, but is inadequate to any preservation effort meant to operate at a scale of millions or even billions of digital objects.) There is no obvious solution to this problem, as neither the donor nor recipient institution can bear the cost of a complete accounting for each and every object in any large ingest. But it does suggest that where there are resources to be invested, investing them in getting the best context possible under the circumstances, at the original moment of transfer, is a very high-value investment.

One of the participants in the AIHT labeled this the GOGI problem: Garbage Out, Garbage In, an inversion of the Garbage In, Garbage Out maxim of the computer science community. Balancing the risk of context loss with the expense of doing a theoretically ideal ingest will be a difficult and continual tradeoff for preserving institutions.

KISS: Experiment, Then Automate

The KISS principle (Keep It Simple, Stupid) is as operative here as in any area of endeavor. Despite the conceptual complexity of the digital preservation problem, we consistently found that we were relying on the simplest tools to handle the archive in the earliest stages, using Unix commands like tar and gzip to pack and unpack collections of files. While these tools were not developed for digital preservation per se, their ubiquity, both as a capability and as a conceptual model, made them attractive, and the simplicity of these tools made them invaluable when trying to understand subtle points of failure, as with the renaming of files.

This parallels a general style of development with networked technologies: the human processes around a certain task (creating and updating the HTML in Web pages, for example) tend to be built piecemeal, using collections of simple tools and a lot of thought and effort. Only after the workflow has been established are tools designed that can automate significant parts of the task (as with Web page creation tools like Dreamweaver).

Much digital preservation is in a similar circumstance, where the processes need to be worked out before they are automated. As a result, we anticipate a reliance on the KISS principle for the next few years, until the community has worked out, in practice, one or more standard workflows for digital preservation. As a design choice, this means favoring simple tools and human judgment in the early period of any digital preservation effort, until enough about the new process is understood to make it a good target for automation.

The Man With One Watch Knows What Time It Is

During the initial transfer, GMU provided three separate representations of the metadata surrounding the archive: the data as it appeared in the file system itself; a copy of an Access database storing the metadata; and a copy of a MySQL database.

No two of these representations showed the same number of files in the collection. As a result, increasing the number of alternative views of what was nominally the same metadata actually increased the work of ingesting the collection while lowering certainty about the "right" answer. The universal response among the participants was to keep the metadata as a secondary source of possible value, but to derive key assumptions by inspecting the archive itself, and deriving things like file sizes and counts from that inspection, rather than from the secondarily reported metadata.

Interestingly, the one piece of metadata that *did* match the archive as transferred was a purpose-built format called TMD, for Transfer Metadata, which was designed by the Library of Congress staff as a kind of digital bill of lading, the simplest possible set of data needed in order for the receiver to tell the sender that the data received was the same as the data sent. This can be accomplished in a brute force way with a single digital digest of a .tar or other aggregated file, but only allows one possible action, which is to resend the entire archive, when the transmission failure may only involve one file out of millions.

By simplifying and canonicalizing the metadata required for simple digital transfer within the AIHT, the Library managed to make the sending of whole archives relatively easy to check, even between institutions with very different assumptions and contexts.

Phase 1: Ingest and Markup

Phase 1 of the AIHT was simply for each participating institution to take possession of an ext3 hard drive containing the GMU archive (the earlier NTFS drives having been abandoned during the setup). Once in possession of this drive, they were to do whatever it took to move that material into one or more local stores, including producing the metadata required for their system.

Though this phase of the work did not require any formal collaboration between the

participants, we discovered that there were a number of issues common to two or more of the participants, often revolving around the difference between the very messy GMU archive and the clean data that many of the systems had been built around.

Requirements Aren't

One of the commonest observations of the participants in the AIHT was that the test was stress-testing the assumptions behind various architectures, sometimes in ways that made previously unexamined risks clearer. The commonest area of risk surfaced was in assumptions around ingest.

Because both the cost and value of any ingested digital content are strongly affected by the quality of the metadata at the original handoff (as per the GOGI principle above), several of the participants designed strong policies and requirements regarding the collection of metadata at the time of ingest, including processes for interviewing the donor, forms to be filled out about the donated objects, and so on.

The design of the AIHT -- bulk ingest of an "as is" archive, with the donating institution already separate from the original contributors of the content -- made many of those policies and requirements impossible to enforce. In particular, ingesting even tens of thousands of objects makes the idea of human examination or conversation about each object unsupportable. At even a minute of examination for each object, ingest of the GMU archive would take half a year of employee time to handle.

Similarly, the temptation to make any given piece of metadata required is really an assertion that a piece of content without that metadata is not worth preserving, or that the institution will expend whatever resources are required to capture any missing but required metadata. In practice, many of these required fields turned out to be unenforceable. The dilemma is that the value of a digital object to any given institution will be more related to the contents of the object than to its metadata, and that any small cost, multiplied across a large number of objects, becomes economically unsupportable. It is obvious that many kinds of contents would be much

easier to preserve with, say, the creator's view of playback environments attached, but it is much less obvious that any piece of content without that metadata will never be worth acquiring. As a result, many of the proposed required fields turned out to be desired fields. The general move here is from a fixed target -- all preserved content will have X fields of metadata -- to a flexible one -- most metadata is optional, but some kinds of metadata are more important than others.

Worse, the size of the GMU archive is a hundred thousand times smaller than what Google indexes today, a corpus that is itself only a small subset of the total digital world. Though the design of the AIHT was in some ways arbitrary, the twin problems of imperfectly or inaccurately described data within large collections are going to be endemic to digital preservation. For a requirement for capturing good metadata at ingest to really work as a requirement, the receiving institution must be able to either force the donor to bear the cost of markup and cleanup of the data, or must be both able and willing to refuse to take noncomplaint material.

Both of these strategies, shifting cost or refusing delivery, are more appealing in theory than in practice. The essential asymmetry of a donation to a preserving institution is that if the donors had the energy and ability to canonicalize the metadata around the donated objects, they would be well equipped to preserve it themselves, the corollary being that donors, especially of bulk collections, are likely to force preserving institutions to face a choice between accepting imperfectly described data and no data at all.

Depending on the value and relevance of the donated material, many institutions are likely to violate their own policies for at least some donations, thus turning them from requirements into preferred options.

As a result, we recommend not spending too much time designing a particularly high or rigid bar for metadata production for donors. Instead, preserving institutions should prepare themselves for a kind of triage: where the unique value of the data is low and

the capabilities of the donor institution is high, insist on the delivery of clean, well-formatted data; where value is high and capabilities are low, accept the archive even knowing that you will be generating both cost and loss in preparing the metadata after ingest; and where value and donor capabilities are both moderate, share the burden of cleaning and marking up the data as equitably as possible.

Metadata Is Worldview

There is not now, and there will never be, a single markup standard for digital content, because metadata is worldview. Metadata is not merely data about an object; it is data about an object in a particular context, created by a particular individual or organization. Since organizations differ in outlook, capabilities and audiences served, the metadata produced by those organizations will necessarily reflect those different contexts.

It is significant that no two proposed systems of metadata designed for use by libraries include the same fields or the same higher-level categories. Even in areas where there is some commonality, as with technical, descriptive and rights metadata, the breakdown of the individual pieces of metadata vary. As a result, the more human judgment encoded in the metadata about a digital object, the harder it will be for that metadata to be losslessly transferred from one regime to another.

Furthermore, even if uniformity of metadata were possible, it would be undesirable, as it would limit the expressiveness of the institution's holding and managing the data. Reducing intellectual expressiveness in the name of increasing interoperability would be a Pyrrhic victory.

There is no solution here, because this is not a problem to be solved but a constraint to be managed. Institutions should understand and plan for differences in worldview, both when they are exporting and importing metadata. Standardization efforts will work best when they concentrate on that subset of metadata that can be reasonably

derived by automated inspection of the digital object itself – file name, size, format, checksum and so on.

Even Small Errors Create Management Problems in Large Archives

The AIHT was designed around an archive that was small in size but varied in format -- many file types, various forms of donation or acquisition and so on. With 60,000 files, a process that correctly handles 99 percent of cases correctly still generates 600 exceptions that, if they require human effort, require hours of work to handle.

We found such exceptions in almost every operation during the AIHT, affecting the creation of file names during ingest, the reading of MIME types during inspection of the objects, assessments of the validity of file encodings and the transformation of the objects from one format to another. The exceptions affected one-off processes, homegrown tools, and commercial and open-source software. No class of operations seemed immune.

The math here is simple and onerous -- even a small percentage of exceptions in operations on a large archive can create a large problem, because it inflates staff costs. And, as noted above, the GMU archive is itself an insignificant fraction of the material that can be preserved. The size and complexity of an archive can easily grow by an order of magnitude. This stands in marked contrast to the difficulty of making tools and processes 10 times better. Yet a 99.9 percent efficient process running on an archive of 6 million items creates exactly the same problems for an institution as a 99 percent/60,000 combination does.

Since the volume of digital material being produced yearly, in both number of objects and total size, continues to grow dramatically, this is another case where the core strategy is triage: Reduce the exceptions it is easy to fix through the improvement of tools. Apply human effort to those exceptions where fixing one error saves either a large number of files or files of great significance. And be prepared to declare some

number of files beyond the economic threshold of preservation.

File Walkers Are an Obvious Toolset

The file is one of the core units in the digital world. Files work as simple containers, (basic text or image files), as encapsulations of complex objects (tar or zip files; PowerPoint or PDF documents that contain smaller binary objects) and as parts of complex objects (a Web site made of html and image files.) Though the conceptual appeal of databases as containers for digital objects is large and growing, the file system is both conceptually simpler (following the KISS principle), and modern file systems are beginning to acquire database-like characteristics. As a result, collections of digital objects are going to be held in file systems for some time as well.

Faced with an archive contained in a file system, and wanting to perform a simple set of operations on every file (reading the extension, inspecting the MIME type, hashing name and contents and so on), several participants developed file walkers that would traverse the file system and apply one or more such checks to the files. Stanford, in particular, invested considerable effort in its Empirical Walker, which uses a plug-in architecture to be able to deploy arbitrary file-checking tools, based on user-defined criteria.

Similarly, JHOVE, Harvard's file-type validation service, has both file-walking capabilities of its own and is callable as a plug-in by other file walkers. JHOVE was one of the two areas of convergent standardization by the participants (METS, noted below, was the other). It was extremely helpful in allowing each of the participants to quickly examine the files in the archive and provided a de facto model for talking about file examination and rendering issues generally. JHOVE is itself an excellent tool and an example of work created in one institution that creates enormous benefits when it spreads to others.

Scale Of Ingest Needs Significant Testing

Scale is a mysterious phenomenon -- processes that work fine at one scale can fail at 10 times that size, and processes that successfully handle a 10-times scale can fail at 100 times. Because we can count on the continued rapid growth in the number of archives, and the number and size of objects contained, predicting the scale at which a given strategy will fail becomes critical.

During the AIHT, there were situations in which ingest of tens of thousands of objects into a particular environment was successful, but once it expanded to include hundreds of thousands, ingest slowed to a crawl. We reached hundreds of thousands of objects because of multiple ingests, which may well happen in the field, since the alternative would be to delete the results of earlier ingests before re-ingesting, a dangerous operation, as it makes every new ingest a single point of failure.

Any working system is likely to be subjected in the real world to a wide variety of pressures on ingest, especially including scale. Institutions offering tools and systems for digital preservation should be careful to explain the scale(s) at which their systems have been tested, and institutions implementing such systems should ideally test them at scales far above their intended daily operation, probably using dummy data, in order to have a sense of when scaling issues are likely to appear.

Phase 2: Export

The Export phase of the AIHT was in many ways the heart of the test. The key issue being tested was whether or not the GMU archive, once ingested into one of the participant's systems, could be easily shared with another participant. The three broad areas of the test were: How difficult would it be to package an entire archive for export? How difficult would it be for a different institution to take in data marked up using someone else's standards? And how much gain, if any, would there be in such sharing over raw ingest of the data.

The descriptions of the difficulties of export and re-import are included in the participants' reports, in Sections 3-6. The consensus view was that there was considerable value in ingesting an already marked-up archive, but that there was considerable variability in that value depending on the similarity or difference in standards and practices between the institutions.

In addition, there were a number of system-level observations made during this phase:

METS and MPEG21 DIDL Are Common

Data formats are set in advance of the acquisition of digital content, sometimes by standards groups (as with ASCII), software vendors (as with .psd or .doc) and sometimes by the file creators themselves, as with custom-created XML formats. Metadata, on the other hand, is created in part by the preserving institution and serves a number of functions, from registering relatively simple items like original file name and size to complex ones like curatorial or contractual limits on playback formats (as with a digital movie only viewable in short segments or maximum size or resolution.)

Creating and handling metadata is in many ways a much thornier set of issues than handling the data itself, since the number, type and complexity of metadata fields is unbounded. As noted above, metadata is worldview, so getting a single comprehensive metadata standard is impossible, for the same reasons that getting a single comprehensive attitude toward preservation among all institutions is impossible.

Three of the four participants (Harvard, Johns Hopkins and Stanford) used the Metadata Encoding and Transmission Standard (METS), and two participants (Johns Hopkins again and Old Dominion) used the Moving Picture Experts Group MPEG-21 Digital Item Declaration Language (DIDL). While the AIHT was a small test, we were heartened to see some commonality in approach among the participants. Any reduction in the number of grammars used to encode metadata, especially locally

invented metadata, which has no independently verifiable aspect, will greatly reduce the cost of sharing data between institutions.

To pass that test, a grammar has to be well-enough defined to merit adoption, easy enough to understand and use to be preferable to homegrown versions and flexible enough to be fitted to local needs and cases. Card cataloging systems such as Dewey and the Library of Congress had to pass these tests as well, and tended only to be adopted when scale of collection and cost of ongoing maintenance broke homegrown systems. We are optimistic that both METS and MPEG-21 pass those tests. If this is correct, then the spread of these standards will tend to make inter-institutional sharing easier, because two institutions using the same metadata standard should be able to quickly identify those areas of metadata conversion that present the hardest problems. In addition, if the number of widely adopted standards remains small, it will improve the market for tools to convert between them, as with the considerable work going on to express MPEG-21 DIDLs in METS and vice-versa.

Future tests and work within NDIIPP involving multiple institutions should always assess the choice of metadata standards among the participating institutions and, where possible, should identify existing work to convert metadata expression among the most widely held of those standards.

Grammar Is Not Enough to Define Expression

As valuable as it may be to have a small number of target metadata standards, it is still not enough to assure semantic interoperability. METS and MPEG21 are grammars, whose adoption will allow two participating institutions to agree on the structure of messages. However, as in natural languages, the number of grammatically correct expressions is vastly larger than the number of sensible and useful expressions.

It is possible to automatically test an assertion in METS (or indeed any rigorously defined grammar) for compliance, but it is not possible to test for comprehensibility.

If a metadata scheme includes an element "foobar" with a rigorously specified set of possible values, a parser will be able to tell you whether any given use of the word "foobar" is compliant or not, but it will not tell you what foobar means.

This is a feature, not a bug; to see wide use, any metadata standard needs to be extensible to local contexts, even understanding that those local contexts will create additional cognitive costs for any future recipient.

Basic grammatical interoperability would be an enormous victory for digital preservation. The goal of NDIIPP in this area should be to reduce, where possible, the number of grammars in use to describe digital data and to maximize overlap or at least ease of translation between commonly used fields, but should *not* be to create a common superset of all possible metadata.

Multiple Expressions Create and Destroy Value

There are multiple ways to describe and store a given piece of digital data. This is true both at the content level, where you can store it in different formats (i.e., canonicalizing all JPEG files to JPEG2000) or different encodings (i.e., storing binary data in Base64), and at the metadata level, where the metadata format, fields and contents are all variable. These multiple expressions both create and destroy value.

They create value by recording content in multiple ways, thus limiting the risk of catastrophic failure, and by allowing different institutions to preserve their own judgment, observations and so on, thus maximizing the amount and heterogeneity of available context.

However, multiple expressions also destroy value, in several ways. At the level of simple bit storage, altered expression of the content itself, whether through format conversion, or the commingling of digital essence with new metadata, will defeat all forms of comparison between various expressions that rely on bit-level comparison, such as the commonly used MD5 digests. Because of the pressure for the use of

simple, well-understood tools, the loss of digest-style comparison will create significant pressure on validation of content held in different preservation regimes, especially when the content is visual or audible in nature and thus not amenable to simple comparison of the rendered product.

At the level of the digital item (as opposed to the bits used to store that item), multiple expressions defeat simple translation of semantics. There are a number of metadata fields such as original file name and format type that will appear in almost any archive, but other fields, reflecting curatorial judgment or notes specific to the local preservation environment, will not be readily translated or used outside their original context.

At the archive level, multiple expressions increase the cost of ingesting archives from other institutions, because of the cognitive costs associated with understanding and translating the various forms of data and metadata in the original archive.

There is no perfect solution here; the ideal number of forms of expression is more than one, to allow for some variability, but less than one per institution, so that some interoperability is possible. The role of NDIIPP here should be to highlight both the advantages and risks of multiple forms of expressions, and to help to find or create tools and techniques for translating between those various forms.

Data Complexity Is Growing Faster Than Computational Power

Moore's Law, a prediction that the density of transistors on a chip will double every 18 months or so, has been operative since the mid-'60s, and has given us the benefit of constantly rising computational performance during that time. Storage density is similarly rising, at a pace that is currently faster than Moore's Law. As a result, it has often seemed that any thorny data management problem will eventually fall to the progress of Moore's law and storage densities.

Sadly, that rosy picture of progress is defeated by combinatorial complexity. Though

having the quality of the hardware double every year and a half or so is remarkable, when data can be combined with other data in various kinds of groupings, and when the metadata that can be attached to those objects can grow in both kind and complexity, the complexity in describing the data will grow exponentially, making mere constant doublings inadequate to the task.

We saw this most clearly during the AIHT in the handling of large XML files. An archive could conceptually contain encodings of all original files in Base64 or a similar encoding, all metadata associated with the original files, full encodings of all format migrations and all subsequent metadata. Because the number of such transformations is unlimited, and because the conceptual scope of an archive could include every digital object held by an institution, there is no theoretical upper limit to the size of an XML file.

As a result, every XML parser that assumes it will build a tree of the data in memory could break. The answer has often been to try to improve the efficiencies of such parsers, which raises the threshold at which they will break, while doing nothing to solve the basic problem -- complexity of files is rising faster than computational power. A tree parser is the wrong kind of parser for very large files; the facts of scale are pushing the world toward handling digital data as streams instead.

More generally, it will be essential, in examining problems of scale, to understand when the potential complexity of any given set of data or handling instructions will swamp current strategies, no matter how long Moore's Law is allowed to run. Scale plus complexity of data sets is likely to push preserving institutions into handling items seriatim within streams, rather than treating the entire collection as a unit to be manipulated all at once.

The Dropbox Was Surprisingly Helpful

As has been noted here several times, the AIHT archive is relatively small. Nevertheless, as archives grow large, the commonest solution to moving large

archives between institutions is to encapsulate the archive as a single large file. Even when this is not conceptually necessary, it is technologically more practical. We discovered that transferring the archive as a collection of files via ftp was more than an order of magnitude slower than transferring it as a single file. We expect single-file transfer to be the norm, whether those files are in some native XML markup, a tar'ed file system, an XML dump of a database or some other format such as Los Alamos National Laboratory's "XML Tape" and Archive.org's .arc format.

As a result, any exchange of large archives among groups of institutions will create difficulties in coordinating transfer, especially when there are multiple expressions of the same content. Even bilateral transfer of content in an agreed-on expression creates some coordination costs because of the requirement for synchronization between sender and receiver.

Faced with this problem, and having magnified it somewhat artificially by having the same archive represented four times, Old Dominion offered to host a "drop box," a server dedicated to only three functions: the uploading of the various archives for inspection; a handful of tools for examining the archives on the drop box, accessible via the Web or ssh; and downloading of any archive chosen for re-ingest.

Though this solution was offered in the spirit of a patch to an existing problem, the drop box proved to be surprisingly helpful in two ways. First, it removed the problem of synchronous transfer, making the transfer of archives closer in design to the store-and-forward pattern of e-mail. The only coordination required for this form of transfer is that the recipient know when the donor has uploaded the content, something which is itself manifest on the drop box and which thus creates no formal communications requirements between parties.

Second, when there are two or more versions of a single archive, it turned out to be very useful to allow receiving institutions to view that content on a remote server,

allowing them to decide which version best fit their needs before beginning the more complicated process of download and ingest.

The pattern was so helpful that it may well make sense for the Library of Congress or some other institution or consortium to create and host a similar drop box capability for the asynchronous transfer of files. Since any reduction in the cost of sharing content will enable an increase in sharing, the drop boxes' ability to reduce required coordination and synchronization costs may be a useful service to offer to the community.

Phase 3: Migration of Format

The Phase 3 test focused on migration of format. Maintaining digital materials for a short term is relatively simple, as the most serious of the playback issues -- altered software, operating systems and hardware for rendering files – do not appear in the short term. Over the long haul, however, merely keeping the same content in the same format actually *increases* the risk of loss, as the continued alteration of the playback environment may make the content unrenderable even though the underlying bits have been perfectly preserved.

The migration test was designed to simulate this problem, by assuming that each participant would choose one file format that the institution would assume had become a preservation risk for unspecified reasons and would take steps to identify and modify all files in the archive stored in the chosen format.

In a real case of format migration, the context for the migration would be clear (several such are possible, from legal encumbrance to lack of supporting software or hardware). Here, the test took place without the sort of real-world context we had during the other two phases. As a result, it was the most mechanical of the three phases and involved selecting a format type for management, selecting a target format to migrate to and designing and running a migration process. For complex data types, such as Web pages, there will be additional complications, such as upgrading all image tags embedded in

Web pages when there is an upgrade from, e.g., GIF to some other graphics format. We did not address those issues, as we were principally concerned with issues of file selection and migration process, precursors to tackling harder issues regarding complex file types.

Because Phase 3 was the test involving the least commonality among the participants, as it involved different content selections and no formal sharing, the differences in choices and approaches meant fewer system-level observations. However, even here, there were some observations that spanned two or more of the participants' efforts:

Playback Drift: The Silent Killer

Many technologists, regarding the problem of digital preservation, begin thinking about the related issue of long-term bit preservation: how, given a string of binary data, can you guarantee its preservation for a hundred years?

Long-term bit storage is a difficult and interesting problem, but it is not the core of digital preservation. We have many examples of perfectly stored but difficult-to-read bits today, at time horizons of far less than a century, such as GIS data commingled with proprietary and undocumented applications written in FORTRAN, which will not run natively on modern machine architectures.

This is a nested set of issues: what format is the data written in? What applications can understand or interpret that format? What operating systems can run those applications? What hardware can run those operating systems? Depending on how far in the future you want to project, one can even imagine asking questions like, What sorts of energy can power that hardware?

This bundle of issues could be labeled “playback drift,” the tendency of a fixed set of binary data to stop functioning or being interpreted in the expected or hoped-for manner, because the complex ecosystem of applications, operating systems and hardware changes, even though the data may be stored perfectly over decades. (In

Internet lore, this problem is called “bit rot,” a humorous and notional condition in which digital data is said to decay with time, even though it is the playback ecosystem that is actually changing.) Indeed, the better long-term bit preservation becomes, the greater the danger of playback drift.

Many of the thorniest issues in digital preservation are affected in some way by format drift, from questions of provenance (Is an updated file "the same" when used as evidence in a court case?) to copyright (Is format conversion a violation of the DMCA?) And because playback drift is really a complex of problems, there is no single strategy that will work in all cases. It is critical, however, for institutions that want to preserve data beyond the timeline of a few years to factor playback risk into their calculations.

Risk Assessment by Format Is a Key Function

One key question tied to playback drift is how much risk is created by the original format of the data? The Stanford team began work on a Format Scoring Matrix, ranking various formats based on criteria like transparency, number of external dependencies and breadth of adoption. Widely adopted, visible and open standards, such as standard e-mail archives in ASCII text, are low preservation risks, at least where playback drift is concerned. Proprietary standards supported by a single vendor (or, worse, by a former vendor, now defunct) and which rely on highly specific configurations of other software or hardware are very high preservation risks.

As a response to inevitable operational constraints in terms of budget and staff, a preserving institution will want to be able to sequence their operations to manage playback drift. Content with low preservation risk can wait to be converted, on the assumption that the content will remain easy to use in its current format for some time. Conversion of content with low value can also be postponed, on the assumption that conversion costs may fall (an implicit “real options” calculation) and that the cost-to-loss tradeoff is in any case low. (There is also an opportunity, in an environment with many communicating peers, to inquire whether any other

institutions regard that content as high value. Transferring such content to an institution that cares about its survival may be a better outcome than holding onto it but postponing its conversion.)

Content of high value and high risk, on the other hand, can be identified using the Format Scoring Matrix plus local curatorial judgment, and this content is the obvious target for early conversion into a format with lower risk, or creation of an emulation environment that will be able to play back files in the at-risk format. (The emulation strategy, of course, merely shifts the format migration problem to the emulator software itself.)

More generally, the process of format assessment will be a continuing one and will have to be fit to various institutional policies (i.e., a preference for translation over emulation or vice-versa). But whenever formats are either translated or put in an emulated environment, care must be taken not merely to handle the alteration correctly, but to make the new format or environment more resistant to playback drift than the old format or environment. This process is analogous to handoffs of the content between institutions, where every handoff creates risks to the content, but the first one is of greater-than-average importance.

Large-Scale Operations Require Fault Tolerance

Big data sets are different than small data sets, in part because even fractionally small errors can mount and because complexity rises so quickly. As a result, workflow must be fault-tolerant. As an example, when transforming a large number of files, it cannot be the case that an error in handling one file causes the entire transformation process to stop. Instead, exceptions must be logged separately, while allowing the process to continue. Similarly, batch ingest processes should be assumed to be flaky, and provisions made for logging ingest failures, so that when an ingest fails, the next ingest can pick up where the previous one left off.

The more general principle is that any workflow that assumes 100 percent efficient

processes will be prone to failure, and that both tools and processes should be created with an eye toward discovering and recovering from mistakes, rather than assuming that all such mistakes can be avoided.

Tool Behavior Is Variable

When considering the viability of a piece of data in a particular format, there is a two-by-two matrix of possibilities. The first axis is correctness: The data either does or does not conform with some externally published standard. The second axis is rendering: The data either does or does not render in software intended to play that format.

The sad truth is that all four quadrants of that matrix are occupied. In addition to the unsurprising categories of correct/renders and incorrect/doesn't render, there is data that fails to conform to the standard but renders in software, and data that passes the standard but doesn't render. You can see the latter two categories at work on the Web today, where noncompliant HTML is rendered by browsers designed around Postel's Law ("Be liberal in the data you accept and rigorous in the data you send out.") and where some compliant XHTML pages do not render correctly in some browsers.

Troublingly, the variability extends even to tools intended to do the conformance checking, where tools meant to validate certain formats themselves have variable implementations, failing or even crashing while reading otherwise compliant files. As a result, there is no absolute truth, even in a world of well-defined standards, and institutions will need to determine, on a format-by-format basis, how to define viability -- by format, by playback or both.

Conclusions

Because the AIHT was conceived of as a test that assumed the autonomy of its participants, most of the conclusions from the test are documented in the final reports in Sections 3-6. Many of these conclusions are related to specific institutional assumptions,

practices or goals and will be most relevant to institutions that are taking on similar problems in similar contexts.

There are, however, a few larger conclusions that we believe will be relevant to many institutions undertaking digital preservation as a function or goal.

Preservation Is an Outcome

Preservation is an outcome. When data lasts through a certain period, then it has been preserved; when not, then not. This is true even if material is preserved through inattention -- a disk drive thrown away and retrieved from a dumpster. It is also true if material is not preserved despite the best efforts of a preserving institution, as with the stored but now unreadable LandSat satellite data.

Having an institutional commitment to preservation, and backing that up with good staff and good tools, only raises the likelihood of preservation; it does not guarantee it. Casting digital data to the winds lowers the chance that it will be preserved, but it does not mean it will be automatically be destroyed. Because the efficacy of preservation can only be assessed after the fact, it suggests that using a variety of strategies to preserve, even within a single institution, may well be a better strategy than putting all efforts toward one single preservation system.

Data-Centric Is Better Than Tool- or Process-Centric at Large Scale

Because NDIIPP has always been conceived of as a multi-participant effort, and because we hosted several meetings with interested parties from many different types of institutions, we have never believed that homogenization of institutional methods or goals was either possible or desirable. In particular, we believe that having many different strategies for preservation, ranging from active management to passive spreading of multiple copies, provides the best hedge against unforeseen systemic failure. The bug in Alexandria was a lack of off-site backup.

As a result, we have become convinced that data-centric strategies for shared effort are far more scalable than either tool- or environment-centric strategies. A data-centric strategy assumes that the interaction between institutions will mainly be in the passing of a bundle of data from one place to another -- that data will leave its original context and be interpreted in the new context of the receiving institution. Specifying the markup of the data itself removes the need for there to be identical tools held by sender and receiver, and the need to have sender and receiver have the same processes in place for handling data.

By focusing standardization efforts on data, and allowing tools and processes to grow in varied ways around that data, we believe we can maximize the spread of content to varied environments while minimizing the cost of doing so. We also believe that that strategy will be more feasible in the short run, because of cost, and better in the long run, because of variety of strategy, than trying to get all the potential partners in a loose network of preservation to converge on either particular technologies or practices.

Shared Effort Is Possible

The essential, and most positive outcome of the AIHT is that it is possible to share effort between interested parties without creating either enormous costs or homogenizing pressures.

It is easy to see how a group of libraries or archives that have some preexisting association, or some standardized set of methods of tools, will be able to share preservation efforts. Spontaneous association, involving two institutions with different tools and no existing relationship, is much harder. The AIHT suggests that by adopting data-centric rather than tool-centric coordination of efforts, institutions will be able to share digital collections without having to converge their assumptions, tools or goals.

There will always be a place for highly standardized methods and processes enforced among a group of tightly coordinated institutions, but as we have learned form any number of systems operating at Internet scale, the higher the requirements for coordination and standardization between two parties, the greater the cost of maintaining the overall system. Any set of standards that lets data be transferred without requiring specific tools or workflow to be in place at the receiving end will tend to maximize the possibility of sharing among an otherwise uncoordinated collection of institutions.

We believe that such data-centric sharing is not only possible and desirable, not only because lowered costs mean greater number of participants, but also because the result of such data-centric sharing -- that many actors with different preservation strategies will hold the same content -- creates robustness against the possibility of systemic failure as a side-effect.

Real Options

When the goal is preservation, it is tempting to subject digital materials to a binary choice: worth preserving, or not worth preserving. There is a third option, however, which is to postpone making the choice. Because of the "small errors in large archives" issues, detailed above, archives will be pushed to declare some digital materials to be not worth preserving.

However, as we can attest firsthand, many tools for describing and handling digital data are still in their early versions. As a result, there will often be value in keeping digital content without preserving it, which is to say ensuring the continued integrity of binary data without expending the effort to ingest, markup and playback such content. This is exercising what financial theorists call a "real option," which is to say an option to delay a choice while waiting for more information or more favorable conditions in which to exercise that choice.

Whenever the cost of merely storing content is small relative to full-scale ingest; whenever the putative value of such materials is high, even if they are too expensive to ingest directly today; and whenever tools can be expected to arise that make format conversion or interpretation simpler or cheaper in the future, the real option of keeping content without taking on the institutional burden or preserving it will become increasingly attractive.

Indeed, the falling cost of storage and the rising number of files held by the average individual is already pushing much of the world to a real options attitude in their personal data management strategies, because deleting files is now a more costly operation than saving them, because the time spent on deletion is more valuable than the storage freed up by doing so. Similarly, many institutions may find the real option afforded by falling storage costs allows them to choose between preserving and merely storing content, a choice that will greatly increase their ability to postpone decisions to delete content they (or another institution) may want later.

Metaphor Matters

Though the AIHT was highly restricted in scope, to create an observable test case, the conversations among the participants were not restricted and ranged over a wide variety of issues. One of the most open-ended and potentially vital was around the question of metaphor.

Digital data is different than analog data, for many reasons. The separation of intellectual contents from any physical container is barely a century old, counted from the invention of the telephone; the ability of digital data to be perfectly copied among various storage media at vanishingly small cost is less than 50 years old; and the widespread use of those capabilities by the general public is only barely in its second decade.

As a result, many of the skills needed to preserve analog data, where inaction can be quite a good a preservation strategy, are either neutral or actively harmful in the case

of digital data, where, for example, inaction greatly increases the risk of loss of playback drift. As a result, the way preserving institutions view themselves and their task will have an effect on the strategies they pursue.

In our conversations during the AIHT, three metaphors for digital preservation arose. The first was the architectural/archeological metaphor, where the construction of an archive is likened to the creation of some hardened object, such as a building, which then survives on its own and little changed. Thought of in this way, the rendering of data in a usable form has to do with limiting ecosystem dependencies and making the format as interpretable as possible, but assumes an otherwise relatively quiescent strategy for the actual holding and management of the digital content.

The second metaphor was religious and organizational, with the principal imperative in preservation being the creation and maintenance of an institution that will survive and will preserve the contents it holds as part of that survival. Here, the rendering of data in a usable form in the future has to do with the success of an institution in surviving, in continually renewing the goals related to digital preservation and in having the ways and means to effect that preservation.

The third metaphor was biological and regarded any given piece of content as part of a digital ecosystem too large and varied to control, but one that afforded a number of low-cost means of preservation. In particular, content that is both interesting and freely shareable tends to be preserved as a side effect of continued use. For example, the famous "Exploding Whale" video exists in many thousands of independent copies on PCs and servers around the globe, because it has the right characteristics for survival in the ecosystem of the Internet. Here, the role of an institution is to help catalyze the spread of so many copies that even if each copy is on an unreliable machine, the sum of probabilities tends strongly toward preservation.

These are just metaphors, of course, and not literal design prescriptions, but we did observe that they affected the way an institution conceived of and executed its role. In

future work, it may be illuminating to document not just the tools and methods an institution uses for digital preservation, but to document its mission, self-conceived, and the metaphors it uses to explain that mission to itself and its partners.

It is obvious that an institution working to package data designed to outlive the preparing institution itself will approach its work very differently than an institution preparing to make preservation covalent with its own survival. What is less obvious is which of those strategies will lead to better outcomes in which kinds of situations. But in keeping with the principle of heterogeneity, it seems likely that maximizing the number of metaphorical approaches to preservation will minimize the risk of a systematic failure.

We Need Data to Be Born Archival

In one of the early meetings on the subject of digital preservation, Stewart Brand suggested that we needed to move the entire system of digital production from material being "born digital" to being "born archival," which is to say created with the attendant metadata and format definitions attached or readily discoverable. Since that meeting, there have been conflicting movements in that arena, with considerable work going on to make file formats harder to decode or convert, for a variety of content from music and movies to PDF files. At the same time, more Web publishing is using RSS, an XML-based syndication format, which allows the content to be accompanied by a few simple fields of metadata, and a number of efforts are under way to specify open formats for standard business documents such as spreadsheets and presentations.

It is not clear whether one of these two efforts will come to dominate, or if the digital world will go to a bimodal distribution of content, with some content becoming increasingly transparent while other content becomes increasingly opaque. It is not clear legally, economically, technologically or politically how the preservation community can affect these issues. What is clear, from a preservation standpoint, is that it is easier to preserve transparent content than opaque content, and that where

content is otherwise unencumbered by business rules or goals that make opacity required, any effort to produce or translate content in a transparent format will lower the cost of all subsequent required actions for preservation.

Section 2 described the issues that arose in common among two or more architectures during the AIHT. The bulk of the project, however, was specific to the individual institutions, where the real work of ingesting and handling the GMU archive happened. Sections 3-6 detail, from the perspective of each of the participants, what they expected from the test, how they prepared for it, what activities they undertook and what observations they made.

Section 3: Report from Harvard

http://www.digitalpreservation.gov/technical/aiht-harvard-final-report.pdf

Section 4: Report from Johns Hopkins

http://www.digitalpreservation.gov/technical/aiht-jhu-final-report.pdf

Appendices available at http://www.digitalpreservation.gov/technical/aiht.html

Section 5: Report from Old Dominion

http://www.digitalpreservation.gov/technical/aiht-odu-final-report.pdf

Section 6: Report from Stanford

http://www.digitalpreservation.gov/technical/aiht-stanford-final-report.pdf

Appendices available at http://www.digitalpreservation.gov/technical/aiht.html

Appendix: Description of the GMU 9/11 Digital Archive

Beginning in the mid-1990s, documents and artifacts of cultural heritage have been increasingly embodied in digital form. Recognizing the wealth of digital information that was not being collected through traditional library channels, many institutions and libraries began programs to collect these bodies of digital heritage. Among these were the American Social History Project/Center for Media and Learning at the City University of New York Graduate Center and the Center for History and New Media at George Mason University, who created the 9/11 Digital Archive (http://www.911digitalarchive.org/) to collect documents and materials about the American experience of the September 12, 2001 attacks on the World Trade Center and Pentagon. In 2003 GMU donated the 9/11 Digital Archive to the Library of Congress with the intention that the Library would sustain it and make it available to researchers long after the collectors at GMU were able to maintain it themselves.

Content Donors

The version of the 9/11 Digital Archive used for the AIHT included contributions from 170 different donors. These donors were private citizens, private organizations, and government institutions. Among the biggest contributors were the Smithsonian Institution, the Department of Justice, and the National Guard Bureau. In addition to the large-volume institutional donors, a large number of stories and images were collected from private citizens through the 9/11 Digital Archive interface.

Type and Volume of Content

The 9/11 Digital Archive contained approximately 57,000 objects that could be classified into seven main categories: still images (IMAGE), web pages (HTML), PDF and "post script" (PDF/PS), text-based documents (TEXT), audio materials (AUDIO), and video materials (VIDEO), and unidentified files (OTHER). See Table 1 below for a breakdown by type, file count and file size for the Archive contents.

Table 1. *Content Types Breakdown by File Count, File Size*

Type	Files	% Total #	Total (Mb)	% Total Size	Avg/File (Kb)	File Extensions Included
IMAGE	15,688	27.3%	5,966	50.5%	389	JPG, TIF, GIF, BMP, OCTET-STREAM
HTML	16,948	29.5%	164	1.4%	10	HTM, ASF, ASP, PHP, PL, CSS, CGI, JS, ADP, JSP
PDF/PS	1,872	3.3%	646	5.5%	353	PDF, PSD
TEXT	20,263	35.3%	98	0.8%	5	TXT, DOC, DAT, EML
AUDIO	2,229	3.9%	2,224	18.8%	1,022	WAV, MP3, AIFF, WMA, CDDA, MPA, MPGA
VIDEO	173	0.3%	2,409	20.4%	14,260	MOV, MPG, AVI, SWF, WMV, QT, RM
OTHER	277	0.5%	296	2.5%	1,093	
TOTAL	57,450	100.0%	11,803	100.0%	210	

Figures 2 and 3 illustrate the distribution of objects according to file size (Figure 2) and number of files by content type (Figure 3).

Figure 2. *Total Size by Content Types*

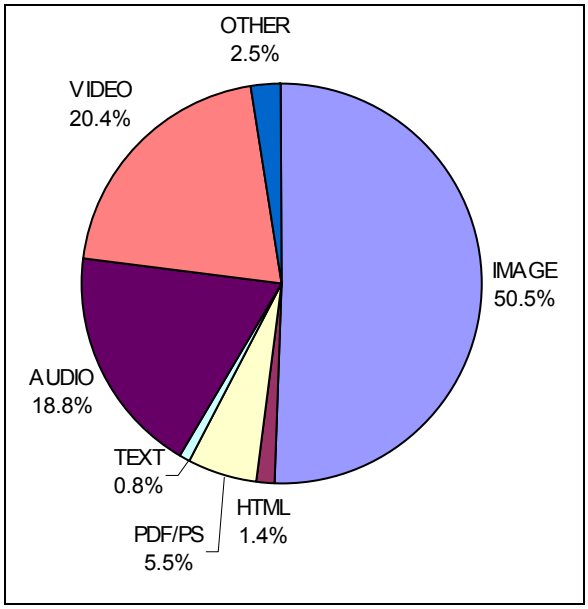


Figure 3. *File Count by Content Types*