Digital Repository Planning and Policy

Lee Mandell and Sue Kriegsman
Harvard University
Cambridge, Massachusetts, USA

Abstract

There is a lot more that goes into a digital repository for preservation than just the technology. Preservation policy is also a key factor in the success of a digital repository. This paper is an in-depth look at Harvard University Library's digital repository and developing preservation plan. The presentation will identify personnel and skill sets, tasks, hardware, software, and networking infrastructure that go into the repository.

In the Beginning…

In 1998 Harvard began discussions regarding a storage repository and then started development on the Digital Repository Service in 2000 as part of the infrastructure of the Library Digital Initiative. The DRS provides Harvard affiliated owners of digital material with a storage and retrieval system for their collections. The DRS and facilities include:

- an electronic storage facility within which the digital objects created or purchased by Harvard agencies reside,
- management of administrative, technical, and structural metadata associated with stored objects,
- preservation policies and procedures to ensure the continued usability of stored objects, and
- a set of delivery services and access management.

The DRS is managed storage and does not accept anything outside of its prescribed, and limited, set of authorized formats. Currently, the DRS is not an institutional repository and only accepts "library like" materials and does not store descriptive metadata. The long term objective of the DRS is to have all materials verified and checked for integrity as part of submission to the repository in order to help simplify ongoing preservation efforts. Once digital objects are stored in the DRS they can be discovered through known applications which include, but are not limited to, 4 of Harvard's public library catalogs.

The goal of the DRS is to have a controlled and cost effective way to preserve digital objects over a long period of time. In order to achieve these goals there are many technology and policy issues that must be addressed.

The Technology

The most obvious items to be reviewed in the creation of the DRS were the technology components. The DRS architecture involves a database, a back-up system, integrity checking, delivery services, naming services, access management, structural and technical metadata, a metadata maintenance system, batch and interactive data loading, a network, and last but not least, digital objects to deposit.

The database itself has already undergone an original implementation and one re-implementation. There is a second re-(re?)-implementation currently being planned for early 2007. So since it's inception in 1998 there will be 3 database versions of the system over 8 years. That's a new database configuration a little over every 2.5 years. Or if we're counting from the time the first version was built, rather than just conceived, then it's a new database a little under every 2.5 years. There isn't really a way to predict if this trend will continue or if it is just a product of the newness of having a storage repository and a steep learning curve.

The DRS is comprised of many components with ORACLE as its database, Tomcat as its application server, and is connected to a Storage Area Network. There is a requirement for high availability for the database and objects so there are failover systems for both. Related software also includes integrity checking of the database as well as the objects. The DRS is starting to work with JHOVE as part of the ingest process. Then there is administrative oversight of the whole system which has to include monitoring for technology obsolescence such as knowing when the current version of ORACLE will no longer be supported and planning for upgrades. The oversight also has to include monitoring of the storage systems as well as the production services.

For each implementation of the database there are tangential components that were also built and have to be maintained. Many of these services are independent of the DRS but are designed to interact with it either exclusively or as one of many systems. The highest priority system is the Name Resolution Server (NRS). This system resolves persistent identifiers to file names to help ensure that as digital objects change location over time the links to the objects will not become lost in space or return a "404 file not found" error message. The NRS resolves persistent identifiers to objects in the DRS but is not exclusive to working with items stored in the DRS. Another system that operates independently of the DRS is the Access Management System (AMS). The AMS controls access to files through a PIN server run by the University. All files stored in the DRS have controlled access though AMS. The

University PIN server was in place but AMS had to be designed to interact with it and be useful for providing access to objects in the DRS at the item level or a larger overarching defined level.

A number of delivery services were designed specifically to work with the DRS. These are the "D.S.s": The Image Delivery Service (IDS) brings images stored in the DRS to a browser, the Streaming Delivery Service (SDS) brings audio files, and video in the future, from the DRS to a user, the Page Delivery Service (PDS) allows a user to navigate through the pages of a logical object, and the most recent addition to the family – the Large Image Delivery Service (LIDS) works with JPEG 2000 image files stored in the DRS. This whole clan of services was designed and built outside of the scope of the original Digital Repository Service and new peripheral services can be added as the demands of the object types stored in the DRS increase. Although these could be considered enhancements to the DRS they are actually independent systems, but they are ones critical to the functionality of the DRS itself.

Each of these items required a specific set of skills and expertise to implement the technology. Someone with the vision to see how each component would integrate with the others was also necessary. Many people had to be involved with the creation of the DRS technology: database designers and developers, system administrators, interface designers, application developers, outside vendors for storage hardware, and database consultants. More than a dozen people were involved when all totaled.

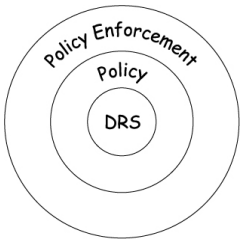
The Policy

Of course, there were still things that were forgotten or simply some things that were reluctantly acknowledged and then conveniently not faced during planning and implementation. Probably the biggest concept that is still debated and dodged is defining what is really meant by a "digital object." There is an implicit notion of a "logical" digital object through the use of the relationship mechanism, but this is not really formally recognized in terms of an abstract data model. Although storing "digital objects" is really at the center of the whole idea of Harvard's Digital Repository Service, it has been very difficult to nail down the definition of "digital object."

Support for collections management is another topic that has been debated and not yet settled. Object owners generally like to have control over whole sets of items in addition to control at the item level. The DRS is supposed to be used for archival storage of digital objects and not as a collection management system so where those two concepts intersect, or collide, is a matter of perspective, preference, and policy.

The technology alone only creates a place to store data and does not make the system a long-term preservation tool because the key difference between bit storage and preservation is creating policy and then enforcing it.

Policy decisions drive the technical architecture and also shape the preservation aspects of the system. Then it is the technology combined with people enforcing the policy that actually makes a preservation based system and not just a data warehouse. Policy dictates what is allowed into the system, what you agree to do with the materials into the future, storage obligations, and a preservation strategy.



It turned out that creating the policies related to the DRS were actually as challenging as designing and implementing the DRS architecture. The first round of policy wasn't actually recognized as policy because it was the system's functional requirements. Nonetheless it required the work of several IT managers, preservation specialists, project managers, collection managers of analog collections, systems administrators, and a development team. Even after the DRS was up and running, there was never really a time when this group was disbanded. 5 years after the implementation of DRS there is still as much discussion and planning going on as there was in the initial design of the system.

What format types will be accepted into the system really became more a matter of policy than technology even if technology was originally driving the policy decision. For example, the original instance of the DRS did not accept audio files. At first there was not a demand from the users to store audio files but the request was eventually made. It was some time before specific audio formats (AIFF and BWF, among others) were added onto the list of accepted formats for the DRS because development time had to be allocated to make the change to the system. Technologically it would have been possible for the DRS to accept audio files when the request was first made but it became policy to wait until the appropriate development and policy work could be done on the system to ensure that the files could be stored and maintained over time. Currently there is starting to be rumor that users may want to store video files in the DRS but it is currently policy that video files will not be accepted until research and development can be put into working with these additional formats.

Along with different format types accepted into the system there are policies about what levels of preservation support can be offered to each format type. All formats are not treated equally. Some formats, such as TIFF and JPEG, have more information available about them than some proprietary formats, such as MrSID or PhotoCD. Harvard made a decision to offer 3 levels of preservation based on the format type:

- *Preferred* – those formats that are most amenable to long-term preservation,
- *Acceptable* – those formats that appear to be susceptible to preservation efforts,

- *Deprecated* – those formats for which only bit-level preservation can be guarantied.

Again, these are policy decisions and not solely based on available technology.

Another policy is the granularity of access control for delivery and administrative services. The Access Management Service was put in place to control access to the objects stored in the DRS but it is also used to control what level of access is available. A curator of a collection may have access to thousands of digital images but temporary project staff may only be given access to a small subset of those same objects. It wasn't enough to just control who was going in and out of DRS but what they could see or download also became a policy decision.

Keeping a curator or object owner involved in the lifecycle of a digital object snuck up as a policy to be addressed and this is still an ongoing discussion. It's not enough for an object to be placed in the DRS and then left behind by the owner and ignored. There has been recent discussion that upon deposit to the DRS, an object owner should sign a submission agreement that includes responsibility for ongoing curatorial attention to the digital object similar to the attention that is paid to a traditional physical object.

Timing became an issue. Both what is acceptable down time for the system, how frequently back-ups will be preformed, as well as how long it takes to ingest objects into the system. The system down time was addressed early in the development and seemed to be an obvious discussion to have. The estimated ingest rate only crept up over time as more objects were deposited more frequently. This policy decision of how long it takes to ingest objects actually drives some technology decisions about what hardware and software configurations will meet the user needs.

Mmm, metadata, the M word. When building the repository it wasn't possible to have a design discussion without raising the issue of what metadata would be stored in the repository and what would be stored outside the repository and the relationships between the data. Harvard decided to store technical, administrative, and structural metadata in the DRS but descriptive metadata is stored outside the system. From a technical perspective this seems like an obvious solution but from an object owner or curatorial perspective it becomes difficult to separate the object from its description. This distinction does make the DRS a storage facility and not a collection management system much to the chagrin of some object owners.

It's also not possible to design a system without talking about who will pay for it. The initial cost of development and implementation was covered by a grant from the University but it was decided that some ongoing costs would be passed on to the users. Currently a subsidized cost for incremental storage is billed to the users at a rate of $5 a gig/per year. Ongoing maintenance and production costs continue to be covered under general library operating costs and central infrastructure. Again, these were policy decisions that surround the actual technical issues.

Once policy is in place it had to be followed up by documentation. The DRS documentation falls into two basic categories: internal documents and external documents. The internal documents cover operational maintenance of the services, disk storage and the database. The internal documents also include administrative issues such as owner, depositor, and maintainer registration as well as billing instructions. External documentation is designed to be consumed by the system users rather than the system designers. This type of documentation frequently has to walk a user through step by step instructions about how to use the DRS or its peripheral systems. There is documentation for the Name Resolution Service, how to deposit objects as single items or in batches, recording metadata, and how to use the web administrative system interface. The documentation also includes any and all forms that have to be completed in order to gain access and permission to the systems.

The reason for the ongoing work on the systems and policies is the nature of digital preservation. It's a new landscape and constantly changing. Trying to keep up with digital preservation is similar to standing on a sand dune and having the bottom shift out from under you at unpredictable times.

The People

There are 13 major roles that are represented at some point during the policy planning:

- Administrator
- Conservator
- Content Provider
- Conversion Specialist
- Financial Officer
- Funder
- Human Resources
- Information Technologist
- Legal Council
- Metadata Analyst
- Project Manager
- Public Services Staff
- User

Any number of people can assume the responsibilities of these roles as long as each perspective is fully represented. The tasks that need to be accomplished to design, build, and create policy for the repository can then be divided among the roles. It is rare that there is a one to one relationship between tasks and roles. For example, deciding what can go into the repository involves a user to state what they want, a content provider to furnish materials, a public services representative to interpret how the user will need to receive materials, a conversion specialist to determine what can be created based on the format of the materials identified by the user, legal council to state what can be converted and stored under the law, a metadata analyst to describe the descriptive, structural and administrative data about the object, IT folks to design a repository that can store and provide access to the

digital object and appropriate metadata, and a project manager to make sure everyone is communicating and staying on track. Not to mention the funder to dole out the money for the project, the financial officer who will control the cost of the project and set fiscal boundaries, and the administrator to provide institutional blessings and support. That's 11 out of the 13 different roles just to decide what can go into the repository.

Each one of these activities involves people time and therefore expenses above and beyond the actual cost of the repository itself. The policy and planning are a secondary cost of creating a digital repository but just as important as the repository.

Certainly an unanticipated staff expense in terms of time and effort has been user support. Every step of the way to set up an individual or commercial depositor, or object owner, has taken a significant amount of time. It hasn't just been a matter of having a robust system with many ways to interact, but there has also been a learning curve about what it means to be part of digital repository. Each user has to learn their own roles and responsibilities and how to work within the system and, of course, the policies.

Conclusion

The staff involved with the design and creation of the system is also in a constant state of learning. There are obvious things to look out for like new trends in digital repositories, other institutional projects, as well as new software and hardware options. As the DRS has become more widely used there is also the challenge of learning how to predict how much new data will be coming into the system each quarter and allocating storage space as necessary.

Harvard University Library is realizing that a digital repository is much more than hardware and software. Policy, and more importantly the enforcement of policy, is key to a successful long term preservation repository. Ongoing planning and re-evaluation have to be part of the system itself. Bringing so many different voices into the planning and development process might seem overwhelming but it is a critical part of ensuring all of the important issues are addressed.

Biographies

Lee Mandell is a programmer/analyst for the Harvard University Library, focusing on systems for visual and archival collections. He received a B.S. in Computer Science from Northeastern University in 1987. He has spent 16 years working with museums and libraries focusing on visual and archival collections management systems. He also spent 7 years working in the pre-press industry helping develop image manipulation tools as well as managing digital work flows. Lee is also a visual artist focusing on photography and sculpture.

Sue Kriegsman is the Digital Library Projects Manager at Harvard University Library where she works with librarians, curators, and archivists as part of the Library Digital Initiative. Before coming to Harvard, she was the Operations Coordinator for the Colorado Digitization Program and worked on the photo digitization project at Denver Public Library. Sue received a B.A. from Alfred University in 1992 and a M.L.S from Simmons College in 1996. She is active with the Society of American Archivists and the immediate past Chair for the Visual Materials Section.