Against the Grain

Volume 21 | Issue 2 Article 8

April 2009

The PeDALS Project

Richard Pearce-Moses
Arizona State Library, rpm@lib.az.us

Recommended Citation
Pearce-Moses, Richard (2009) "The PeDALS Project," *Against the Grain*: Vol. 21: Iss. 2, Article 8.
DOI: https://doi.org/10.7771/2380-176X.2551

Unchartered Territory ...
from page 38

sities had content identified from the start. **UCSB** ingested the geospatial content from the **California Spatial Information Library** (**CASIL**), which included scanned topographic maps, **LANDSAT** imagery of the state of California, thematic data layers including transportation, boundaries, elevation, farming, and structures. **Stanford** accessioned the **David Rumsey Collection** of 18th and 19th century scanned historical maps and the output (maps and field notebooks) of the **Stanford Geological Survey**. The collections continue to grow rapidly with **UCSB** acquiring the **Citipix** aerial imagery collection of 65 metropolitan areas across the United States with over half a million images. **Stanford** has collected high resolution imagery of the San Francisco Bay Area, elevation data, data layers from the *National Atlas*, coastline data, and scanned aeronautical charts.

One of the current challenges we at **Stanford** are addressing is setting up a structured workflow for the data life cycle. For example, we acquired imagery and elevation data from the **United States Geological Survey's EROS Data Center**. It was delivered on a hard drive. The data then had to be reliably duplicated on another storage medium in case the hard drive failed. Metadata was not included and so had to be pulled from the **USGS National Map Seamless Server**. Now that the metadata and the content are in place, decisions have to be made about how the content will be stored in the archive — as a whole collection or in its individual parts. The data and imagery then must also be brought into the library workflow for patron use with cataloging, display options, and the ability to download the files of interest. There are many pieces to the puzzle with potential failure points in numerous spots along the way; our approach is piecemeal and not yet fully formed. The goal, by the end of the agreement with the **Library of Congress** (August 2009), is to have a comprehensive workflow for our digital acquisitions that is as seamless as the process for our paper-based materials.

Finally, a format registry is being created as a joint effort by both universities to maintain technical information about the formats being archived. The registry will house specifications, standards, white papers, and ancillary information about the formats in order to increase the likelihood that they will be understood and usable in the future. It has been a complicated process to decide exactly what should be kept, where it should be housed, and when to say enough is enough in terms of the amount of information collected. We have been watching the developments of similar projects at **Harvard's Global Digital Format Registry**[2] and the **United Kingdom National Archives' PRONOM**[3] projects as we would eventually like to pool our registry information.

Conclusion

The work on the **NGDA** project has been challenging, interesting, and critical to the success of the geospatial collections at both schools. While it is easy to grab digital content and bring it in house, it is entirely a different matter to make sure that access is provided now and into the future as securely as any book we pull off our shelves. It is our hope that the work we have done to address and resolve some of the issues inherent in geospatial data collection will be of use to others in the field. At our Website, *www.ngda.org*, we have posted the collection development policies, contracts, the **NGDA** interface to view a sample of the collections, articles and publications, tools, and technical architecture specifications.

Endnotes

1. *U.S. House of Representatives Report* 106-1033 Making Omnibus Consolidated and Emergency Supplemental Appropriations for Fiscal Year 2001. *http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=106_cong_reports&docid=f:hr1033.106.pdf* (Accessed March 18, 2009).
2. **Global Digital Format Registry**. *http://www.gdfr.info/* (Accessed March 23, 2009).
3. **The Technical Registry PRONOM**, *http://www.nationalarchives.gov.uk/PRONOM/Default.aspx* (Accessed March 23, 2009).

The PeDALS Project

by **Richard Pearce-Moses** (Deputy Director for Technology and Information Resources, Arizona State Library, Archives and Public Records, 1700 W Washington, Suite 200, Phoenix, AZ 85007; Phone: 602-926-4035) <rpm@lib.az.us> *http://pedalspreservation.org*

Archives have a number of requirements that distinguish them from other types of repositories. When developing a digital archive, archivists must find practical IT solutions that meet these requirements within the specific context of their repository.

Given the fundamental shift from tangible to virtual materials, archivists have to reconsider all aspects of curating a collection, from selection, through acquisition and processing, to storage and long-term preservation, and use. Currently, no single approach has yet to be widely adopted, so there are no well-established best practices. A number of organizations are building systems, and the different projects are learning from each other.

The **Persistent Digital Archives and Library System** (**PeDALS**) project[1] is a research project that seeks to articulate a curatorial rationale that describes an automated workflow for processing collections of digital archives and publications. The project seeks to learn lessons about how the nature of curation changes in the digital era. The project is led by the **Arizona State Library**, **Archives and Public Records**, with partner state libraries and archives from Florida, South Carolina, New York, and Wisconsin. The project is funded by a grant from the **Library of Congress**, **National Digital Information and Infrastructure Preservation Program** (**NDIIPP**).

This article describes some of the archival requirements for storage in a digital archives system and how **LOCKSS** (for **Lots of Copies Keep Stuff Safe**) meets those needs.

Controlled Access

When starting an archives, possibly the most crucial first step is to identify a secure place to store the records. The archives must be able to control use of the materials so that these valuable materials do not disappear through malice or neglect. The storage space does not have to be ideal. For paper records,[2] it could be a closet, a file cabinet, or small storge container that can be locked to control access. Because paper records are reasonably stable, securing paper records buys significant time. A controlled environment, advanced security, and acid neutral containers can come later. Even unstable paper records can be used for many years if those records are kept in an ordinary office environment and much longer if kept in a carefully controlled environment.

Unfortunately, digital records are not nearly as stable as paper records. The problems of digital preservation are generally well known. The signal on the media is much more fugitive than ink on paper. The life of software and hardware used to render the records is measured in years, not decades. Because of the fragile nature of digital media, archivists do not have time to find new ways to store, preserve, and access electronic records. While secure storage is still a critical first step, preservation must be addressed very quickly.

Longevity

One distinguishing characteristic of archival records is their "ongoing usefulness."[3] As a result, archival records are often described as being permanently valuable. Professional archivists often prefer the phrases "enduring value" or "continuing value," but — to use the vernacular — archives are repositories for records that must be kept for a very long time.

In the recent past, IT has appropriated the term "archives" for electronic data that are seldom used, but must be kept for a period of time before being discarded. These data are

continued on page 42

often kept on tape to reduce the cost of online storage. These systems may be marketed as a means to store records for a long time. However, within IT, ten years is often considered a very long time, hardly the same time frame in an archivist's mind.

Possibly the first roadblock to building a digital archive is to ensure that IT professionals on the project understand what archivists mean by a long time. Examples help. The **National Archives and Records Administration** (**NARA**) holds the *Constitution*, a document that has been in use for centuries. Land records are kept permanently to ensure clear title to deeds. Birth records remain in use for decades, throughout an individual's life, and beyond death for historical and genealogical purposes.

Fixity and Integrity

Another characteristic of archival records is the unchanging nature of the information they contain. The records serve as a reliable voice from the past, and that reliability is based on the stability of their form and content. Records may suffer some degradation over time without seriously affecting their reliability as evidence of the past. Paper may yellow and inks may fade, but the record remains readable.

Demonstrating the integrity of electronic records is more challenging. Current methods use a hash value that can detect a change to an individual bit. Unfortunately, those tests cannot indicate whether the change has a significant impact on the content. A single flipped bit might look like a typo or speck in an image. But if information is encoded as a binary zero (no) or one (yes), a flipped bit could completely reverse meaning.

Archivists must find systems that can spot changes resulting from degradation and correct those errors. Current practice keeps two copies of every file. The system constantly checks for degradation, and replaces a corrupted version with the second — presumably — correct copy.

Preservation of Unique Records

Archival records are also distinguished by the fact that they are typically unique. Loss of a single publication distributed in even a modest run is mitigated by the availability of other copies. If a copy is destroyed in a disaster at one repository, other copies are likely available at other repositories.

Because records have no redundant copy, archivists take exceptional care to protect their holdings. A second copy of records kept as a check against loss of integrity can be stored offsite, eliminating the risk of losing unique copies. The ease of duplicating digital records and transferring them for offsite storage is one of the greatest benefits of digital records over paper records.

Authenticity

Another important characteristic of archives is the need to ensure the authenticity of the records. The archives must be able to demonstrate that the records are what they purport to be, that they are genuine and not fakes or forgeries. As records are often used in litigation, archivists must be able to demonstrate that the records meet the requirements of authenticity outlined in a court's rules of evidence.[4] When used for historical research, the authenticity of the records serves as a foundation for understanding the past and is essential for factual scholarship.[5]

Cost

Even in the best of times, archives seldom have adequate resources. Like libraries, archives are now being asked to do double duty, working parallel in tangible and digital universes, and often with little or no additional funds. Unfortunately, investments in information technology can be very expensive. Some commercial systems designed to meet archival requirements may be prohibitively expensive for even medium and large sized archives.

Why LOCKSS?

Archivists and IT professionals must work together to find solutions that can keep archival records accessible for a very long time. They must ensure that the records' integrity is preserved, that the records are protected against disaster, and that their authenticity can be demonstrated. All this must be done within a limited budget. Unfortunately, because digital preservation is so new, there are no time-tested best practices.

The Arizona State Library, Archives and Public Records is the official archives for the State of Arizona and also serves as the custodian for local governments' archives. The agency also serves as the official depository of state agency publications. The agency has not yet allowed archival records to be deposited in digital format, but it is under increasing pressure to do so. The agency has effectively been forced to accept digital publications, as many of those documents are never printed. Arizona needed a solution.

Over the past several years, staff has taken the first step of creating secure storage for digital records and publications. However, that system failed to address all the archival requirements for a robust digital repository described here. Commercial vendors often failed to understand the particular needs of the system, especially the need to build a system that could support permanent retention. Vendor systems required both a large up-front investment plus significant ongoing costs in personnel for support.

While **LOCKSS** was originally conceived as a system for serial publications, a certain parallelism suggested that the technology might be adapted to archives. Where serials have a publisher, possibly with many titles, and many issues within a title, archival records have a provenance, possibly with many series, and many records within those series. On further investigation, **LOCKSS** clearly addressed the

continued on page 43

against the grain

people profile

Richard Pearce-Moses

Deputy Director for Technology and Information Resources
Arizona State Library, Archives and Public Records
1700 W Washington, Suite 200, Phoenix, AZ 85007
Phone: (602) 926-4035 • <rpm@lib.az.us>

PROFESSIONAL CAREER AND ACTIVITIES: Curator of Photographs, **Department of Archives and Manuscripts**, **Arizona State University Libraries**. Archivist and Automation Coordinator, Heard Museum. President, **Society of American Archivists**, 2005-2006.

PHILOSOPHY: In medio stat virtus.

MOST MEANINGFUL CAREER ACHIEVEMENT: "A Glossary of Archival and Records Terminology" (**Society of American Archivists**, 2005). *http://www.archivists.org/glossary/*.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: Have helped build a framework for the automated processing of archival electronic records.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: The archival profession is under enormous pressure to accept digital records in a wide range of formats, even though archivists know that these media are difficult — maybe impossible — to preserve for long periods of time. I believe the profession is making significant strides in learning how to work with these materials, with the result that in five years I believe the profession and its practitioners may look very different.

The PeDALS Project
from page 42

distinguishing requirements of an archival repository.

The **LOCKSS** team understands the need to keep information resources for a very long time. As a result, they have already been thinking about archival storage system requirements, even if in a different context.

LOCKSS supports automated integrity checking and error correction. The technology required no adaptation to meet an archival repository's need for fixity and integrity. **LOCKSS** was built to support a distributed preservation network by keeping copies in multiple locations. Again, the technology did not need any modification to meet a critical preservation requirement. In fact, **LOCKSS** is outstanding as a preservation system; some commercial systems that keep multiple copies do not offer distributed storage, but keep both copies in a single system. Finally, **LOCKSS** uses a sophisticated polling technique among multiple copies to protect the records from a malicious attempt to replace authentic records with forgeries; this methodology makes it particularly easy to demonstrate the authenticity of the records. Finally, **LOCKSS** is significantly less expensive than any other commercial solution.

Some Concerns

Agency staff had a number of concerns about **LOCKSS** while developing the **PeDALS** architecture. They had lengthy discussions with **LOCKSS** staff about these potential problems. Agency staff also consulted with the **MetaArchive Cooperative**,[6] which had already implemented a distributed preservation network for special collections materials using **LOCKSS**. Through these conversations, agency staff determined that their concerns could be readily addressed.

First, archival collections contain many records, which raised the issue of the capacity of **LOCKSS** as a storage system. **LOCKSS** is built on top of **UNIX**, which can easily accommodate terabytes of digital data. However, the **UNIX** file system has practical limits on the number of files it can address. Given that many archival records are rather small in size, staff was concerned that the repository would reach the file limit long before it would reach storage capacity. The solution was to store collections of records in "super packages." For example, all records in an acquisition would be encapsulated within a single file.

Staff is still concerned about the maximum capacity of a **LOCKSS** system. The time necessary to perform integrity checks on all the files in the system places a practical limit on the size of a **LOCKSS** system. At the moment, **LOCKSS** staff believes maximum capacity to be approximately ten terabytes, assuming relatively low-cost servers. **Arizona State Library and Archives** is investigating the use of more powerful servers to address that issue. Regardless, the cost of a **LOCKSS** system is low enough that it will be possible to implement additional **LOCKSS** systems.

Second, because **PeDALS** will contain records that must be kept confidential by law, the system must be a private network. This requires the system to have multiple **LOCKSS** servers, each with a complete set of records. This differs from the use of **LOCKSS** to store serials, where many different libraries would capture the same serial. Some serials may be captured by dozens of libraries and no library need to create redundant copies. A **PeDALS** system will include seven **LOCKSS** servers distributed across at least three states.

Finally, agency staff was concerned about risks associated with the use of open source software. Where commercial software has a vendor backing the product that can provide product support, open source software typically relies on volunteers. At first glance, open source may appear to be an unreasonable risk for an archival repository. However, some commercial software has been abandoned, and **Linux** has a large and committed development community. Agency staff believes that the level of risk associated with using **LOCKSS** to be acceptable. Although **LOCKSS** does not have the backing of a commercial enterprise or a large open source community, it does have a significant number of organizations willing to support the technology's ongoing support and development.

All told, staff felt that the costs and benefits of **LOCKSS** far outweighed these risks. Since then, staff has considered a more limited use of **LOCKSS** for robust, near line storage for digitized images. In this context, many archival requirements are largely moot because the original paper record is preserved. However, **LOCKSS** offers a robust mechanism to store the digitized image and ensure that the work of digitizing the images is not lost due to failing media or single copies.

Endnotes

1. Information about the project is online at *http://pedalspreservation.org/* (Accessed March 24, 2009).
2. The paper records may be textual or graphic. Further, many other media are similarly stable, such as film and glass. For the sake of simplicity, "paper" will be used throughout to refer to traditional record formats that are reasonably stable over time.
3. **Richard Pearce-Moses**, *A Glossary of Archival and Records Terminology* (Society of American Archivists, 2005), online at *http://www.archivists.org/glossary/* (Accessed March 24, 2009). Archival value is defined as "The ongoing usefulness or significance of records, based on the administrative, legal, fiscal, evidential, or historical information they contain, justifying their continued preservation."
4. The **Federal Rules of Evidence** are available online from the **Cornell University Legal Information Institute** at *http://www.law.cornell.edu/rules/fre/* (Accessed March 24, 2009). Each state has its own rules of evidence, although they often follow the **Federal** rules closely.
5. For an excellent illustration of authenticity in historical research, see **Peter B. Hirtle**, "Archival Authenticity in a Digital Age," in *Authenticity in a Digital Environment* (Council on Library and Information Resources, 2000), p. 8-23. Available online at *http://www.clir.org/pubs/abstract/pub92abst.html* (Accessed March 24, 2009).
6. Information about the project is online at *http://www.metaarchive.org/* (Accessed March 24, 2009).

Rumors
from page 32

The article is about famous deceased authors and the various files and artifacts that they are leaving to libraries in formats that are no longer in use. "'The floppies … are outmoded and damage-prone by today's standards.' [says] **Ms. Morris**, who curates modern books and manuscripts [at **Harvard University's Houghton Library**]." I mean, y'all, I bought a **Kindle 1** in September, 2008, and everyone is now making fun of me for not having a **Kindle 2**. They are calling me "retro." See – "Archiving Writers' Work in the Age of E-Mail," by **Steve Kolowich**. *http://chronicle.com/weekly/v55/i31/31a00102.htm*

And, speaking of deceased authors, saw recently that the **University of Massachusetts W.E. B. Du Bois library** in Amherst is going to post **W.E.B. Du Bois**' documents (estimated at 100,000) online. It is projected that the task will take two years and help from a $200,000 grant from the **Verizon 29thFoundation**, which funds scholarly programs that use technology. **Du Bois** died in 1963. The library got the papers from his widow, **Shirley Graham Du Bois**. The materials (papers, letters, diaries, photographs, speeches, essays, etc.) have been largely inaccessible except to the most dedicated researcher. **Rob Cox** is head of special collections at the **W.E.B. Du Bois** Library. See – "UMass to Post Treasure Trove of Du Bois Documents Online," by **Peter Schworm**, ***The Boston Globe***, April 4, 2009. *http://www.boston.com/news/education/higher/articles/2009/04/04/umass_to_post_treasure_trove_of_du_bois_documents_online/?rss_id=Boston.com+--+Education+news*

And, **in homage to a book**, wanted to tell y'all that one of the most influential chemistry resources in the world has turned 100! Since 1909 **Houben–Weyl** has been used by chemists working in academia and industry. In 1909, **Theodor Weyl** wrote and edited the ***Houben–Weyl Methods of Organic Chemistry*** series. The first edition, consisted of two volumes and covered material published from as early as 1834. In 1913, **Josef Houben** expanded the project. The two German chemists made a significant contribution to the field of chemical information at the commencement of the 20th century. **Weyl** and **Houben** were the first scientists to exhaustively evaluate the organic chemistry literature with regard to its practical application. In order to mark the **Houben–Weyl** centenary, 100 selected articles

continued on page 49