Persistent Digital Archives and Library System:
Final project report to the Library of Congress
April 19, 2012

Executive summary

The Persistent Digital Archives and Library System (PeDALS) research project was funded by the Library of Congress' National Digital Information Infrastructure and Preservation Program as part of its Preserving State Government Information initiative. The project explored the development of a curatorial rationale to support an automated workflow to process collections of digital publications and records, specifically using Microsoft BizTalk Server middleware to manage the collections and rules-based processes for their ingest. PeDALS also examined the practicality of Stanford University's LOCKSS, or Lots of Copies Keeps Stuff Safe, storage networks as an effective and inexpensive method of distributed preservation. In addition to those technical goals, PeDALS worked at building a community of shared practice among its partner states in the hopes that shared software development and best practices would foster a system that could be applied to a variety of repositories. The PeDALS partners are the Alabama Department of Archives and History, the Arizona State Library, Archives and Public Records, the South Carolina Department of Archives and History, the State Archives of Florida, the New Mexico State Records Center and Archives, the New York State Archives, the New York State Library, and the Wisconsin Historical Society. In our report, we have included the processes that worked and those that did not because we believe that we learned much from both. We feel it is important to share these since other institutions may want to implement a PeDALS system and these institutions will benefit from our experiences.

PeDALS has been a four year collaboration. Despite a faltering economy that found all of the partner states reeling under furloughs, budget cuts, layoffs, and outright facility closures, a core group of states and individuals have remained committed to the project. These states now have the basic framework and workflow in place that accomplishes much of what the original project planners had hoped for: an automated method of ingesting collections of digital records into storage, with discovery and access to the original record and an access version of the record more suited for distribution over networks where bandwidth and storage may be an issue. Several of the states have ingested and preserved government records that were at risk of loss.

PeDALS offers an unique solution for the preservation of large record series of permanent government records. Among PeDALS most notable features:

- Created a complete modular, integrated process that validates the curatorial rationale through a flexible automated workflow, based on the Open Archival Information System (OAIS) reference model, for the curation processes of high volume digital records.

- Established and validated pattern-based templating methodologies and mechanisms to accelerate series-specific development processes.

  Since the aim of the overall process is to automate the unification of useful and appropriate metadata with a digital record both for archival storage and potentially discovery and dissemination, only two basic automated workflow processes are necessary. One in which the record is itself self-describing or, more commonly, a

consistent metadata pattern exists, may be exported, or is quickly established. Each series specific solution would then only require the basic processes be customized to account for the differences in both input and appropriate output metadata schemas.

- Established and validated series-specific system processes that allowed the capability for rule-based ingest of records to be developed and deployed by less technically skilled staff.

- Developed several limited digital record transformation components for prototype dissemination, fixity and the capability for future transformations of the data while preserving the record in the original format.

- Created and validated strategies and implementations that identified opportunities to further improve system capability, compatibility and flexibility while further decreasing technical requirements for system and solution development, deployment and operation.

  The initial example of this strategy was the early modular design of the released beta version of the PeDALS PST preprocessor, which evolved through a similar descriptive metadata parsing and extraction tools developed for other specific project purposes. The architectural design of the PeDALS Dissemination Information Package Processor represents implementation of a late stage evolution of the design strategy. Much of its code will serve as the code base for the next version of the PeDALS system. This will represent a shift away from a processing system of two customizable but monolithic workflows heavily dependent on constituent formats and schemas to a highly configurable and scalable system that utilizes a large degree of format independence to provide an ever increasing degree of flexibility and capability in small, manageable and testable atomic increments. This strategy also includes use of advanced solution templating mechanisms that are expected to require only Microsoft BizTalk mapping skills or basic string formatting programming skills for series specific solution development in all but the most complex automated curatorial process scenarios.

Project history

2006-2007

In 2006, the Arizona State Library, Archives and Public Records began to experiment with possible designs for a digital archive. The Library's chief technology officer, Richard Pearce-Moses, identified that the problem of handling digital objects, particularly those that are born digital, was as much a problem of practice as it was a problem of technology. He felt that the traditional methods used by archives still applied in the digital realm – that the primary functions of acquisition, description, and preservation were every bit as relevant to digital objects as they were to documents. The Library's first attempts at developing a digital archive were experimental, with the hope that staff could acquire some practical experience and knowledge that would be useful when developing specifications for a production system.

At the time, the Washington State Archives was developing a system that used Microsoft's BizTalk Server middleware to automate a systematic workflow for processing electronic records. Although this workflow required a significant investment in

programming time, spreading the cost over thousands of records, particularly records that would be transferred to the archives on a recurring basis appeared to make the process cost-effective. Reusing the code for different series of similar record types also would further justify the initial expense. Arizona staff members took a Microsoft course in BizTalk. Based on the knowledge gained from the classes, Washington's successes to that point, and the fact that Microsoft generously offers a substantial software discount to libraries, Pearce-Moses decided that BizTalk was a reasonable solution for the initial exploratory project.

2008

The Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) issued a call for proposals under the Preserving State Government Information initiative. The Arizona State Library, partnered with several other states, submitted a proposal for what would become the Persistent Digital Archives and Library System (PeDALS). On January 7, 2008, The Library of Congress awarded four state government projects grants, including PeDALS with Richard Pearce-Moses as Primary Investigator. The original project partners were the Arizona State Library, Archives and Public Records, the State Library and Archives of Florida, the New York State Archives, the New York State Library, and the Wisconsin Historical Society. The South Carolina Department of Archives and History joined the project in May 2008, paying for much of its own PeDALS hardware and software rather than relying on grant funding.

Pearce-Moses started work with the original partners on three fronts: establishing a digital workflow for archivists that would stay faithful to their professional standards, developing a metadata standard for digital records which would be used to populate the administrative database, and the coding of business rules in BizTalk which was the basis of the automation process. Much of the preliminary workflow design had been done prior to beginning the project.

On Jan. 16, 2008, representatives of the five initial PeDALS partner states met for a project kickoff meeting in Phoenix, Ariz. The first order of business was to form a metadata standard group, which was led by Jonathan Nelson from Wisconsin. The metadata standard group worked through regular conference calls of the group as a whole and by dividing the work among three subgroups; administrative, preservation, and discovery metadata. Bonita Weddle of New York and Bryan Collars of South Carolina contributed significantly to the effort to incorporate the best of other metadata standards, such as PREMIS, MARC and Dublin Core into the PeDALS metadata standard. State representatives worked to create a standard that would meet the varying needs and requirements of all the participating states. This goal was met by limiting mandatory metadata requirements and expanding the number preferred and optional metadata fields. A first draft of a metadata dictionary was completed in May 2008.

In April 2008 David Lindsay, a Microsoft-certified BizTalk instructor who had provided the initial BizTalk training to Arizona staff members, offered recommendations on how best to use BizTalk within the system architecture of PeDALS. The project contracted with Bryan Vincent and Associates to review Lindsay's proposals and to assist in creating a prototype system. Technical staff from Florida, South Carolina, and Wisconsin received BizTalk training in a weeklong class in Columbia, S.C., in September 2008. Afterward, Wisconsin's Dennis Bitterlich and South Carolina's Matt Guzzi began experimenting with what BizTalk could accomplish for PeDALS.

At the project partners meeting in Phoenix , October 20-24, 2008, the PeDALS technical team discussed what system hardware would be needed. The attendees also examined the design and discussed how best to set up the administrative database. Based on those discussions, hardware and software was ordered in November 2008, for the initial PeDALS systems to be built in Florida, New York, and Wisconsin. Each installation included 3 servers, Microsoft's BizTalk software and Microsoft SQL for the database. Seven additional servers from Iron Systems to be used for the LOCKSS (Lots of Copies Keeps Stuff Safe) clusters were also purchased. These servers would store the records for preservation purposes. Tom Lipkis, a LOCKSS systems architect, worked with Pearce-Moses and Guzzi in December 2008 and January 2009 to set up the first working LOCKSS clusters for PeDALS in Arizona and South Carolina.

2009

The technical team met again March 19-20, 2009 in Phoenix for a "code-a-thon," during which they continued work on the database design and BizTalk code. In the spring and summer of 2009, several pieces of the PeDALS system were designed by the project staff. The administrative catalog database was developed primarily by Dennis Bitterlich, in Wisconsin, while working with Pearce-Moses and Sara Muth in Arizona. A web tool to enter data into the administrative catalog and to see the status of ingests while BizTalk was working was written by Alan Nelson and Slawomir Mitak in Florida. Nelson and Mitak collaborated with Muth and Bitterlich, on scripted queries to the database. Pearce-Moses then contracted with Neudesic, a BizTalk development firm with offices in Phoenix that was recommended by Microsoft, to assist in the implementation and configuration of BizTalk. Pearce-Moses and Neudesic identified a set of digitized marriage certificates with accompanying XML metadata files, as the first data set of records. Neudesic worked with the PeDALS tech team through weekly "scrum" conference calls, based on the Agile software development methodology.. However, it was several months before Neudesic was able to produce working code. During this time Neudesic had, on three separate occasions assigned new program developers to the PeDALS contract. This caused large delays in development as each programmer needed to be trained in archives and preservation theory and standards as well as being brought up to speed on system architecture and code already developed. Also in the summer of 2009, Muth found her time as PeDALS project manager being allocated to other Arizona State Library work requirements and, therefore, could no-longer devote time to the project. Pearce-Moses hired Pete Watters in August 2009 to be PeDALS project manager while he focused on project duties.

The first iterative review of the metadata standard was completed in May 2009. Bryan Collars of South Carolina led a committee that began work on revisions and additions, with most of the work being finalized by November 2009. Representatives from Chief Officers of State Library Agencies (COSLA) and the Council of State Archivists (CoSA) were invited to participate as observers in these final discussions. In 2009, those PeDALS metadata fields that had analogs to fields from the more-familiar standards were cross-referenced to make them more accessible to archivists and librarians who were encountering the PeDALS standard for the first time.

In late August, Neudesic demonstrated functional BizTalk code that accepted the series of marriage certificates from the Archival Information Package (AIP) into the LOCKSS system. While the code showed a methodology that came close to the specifications Pearce-Moses had requested, it was tailored specifically for the marriage certificates and as a result

was not easily transferable to other records series, such as the ones that South Carolina and Wisconsin staff were working on processing.

Due to the excessive turnover of contracting staff and the code not working as expected, Pearce-Moses decided to hire a BizTalk developer that could work solely on the PeDALS project. In October 2009, PeDALS hired Brian Schnackel, a .NET developer, to work with Neudesic to create a model that would allow partners to reuse the code with different record series. In September, Neudesic turned the PeDALS project over to a fourth developer, Madhukar Konda. Watters and Schnackel worked closely with the Konda to remove unnecessary code and streamline workflow. They had an efficient BizTalk application written for marriage certificates by the end of October 2009.

In the fall of 2009, PeDALS received a supplemental grant from NDIIPP to continue the development of the PeDALS system. September 2009 also saw the addition of two new PeDALS partners, the New Mexico State Records Center and Archives and the Alabama Department of Archives and History. During 2009 all partner states saw large staff reductions, furloughs, reassignments and increased work loads of remaining staff, and large budget cuts, due to the economic down turn, causing significant impact to all state partners. Florida state budget cuts and severe staff reductions, coupled with a state mandated "no travel" restriction, left the State Archives of Florida unable to continue as an active participant in the PeDALS grant project.

Representatives from all states, with the exception of Florida, met at the "All Partners" meeting in Columbia, S.C., in mid-November and the new marriage certificate code was demonstrated to attendees. Software and hardware was also ordered for the new Alabama and New Mexico partners.

Late in 2009, PeDALS contracted with Neudesic to provide BizTalk training specific to the PeDALS code it had developed. The weeklong course took place in Phoenix, in December 2009, for the technical staff from New Mexico, Alabama, South Carolina, and Arizona. PeDALS contracted with Neudesic to write code for another record series, this time ingesting Microsoft Outlook PST files, the most common email file-type that partner states were being requested to archive by outside agencies. Working closely with both Watters and Schnackel, Neudesic's programmer Konda tackled the project through November and much of December. At the time, Schnackel had combined code from two open source code utilities, PMSEU – personal mail storage extraction utility and PST Walker, to include the best of both. He then added additional coding to create the PeDALS PST parser, that parsed PST files into individual XML records that could work with the BizTalk process. For small test samples that had been curated beforehand, the code worked well. However, when the code was run on large, unstructured PST files provided by Florida, Alabama, and Wisconsin, the open-source based tool failed, and the team was unable to ingest the files with PeDALS.

2010

2010 was a time of great change for the PeDALS partners. Dennis Bitterlich left his position in Wisconsin for Texas early in the year and was replaced by Sarah Grimm in mid-March. He continued to contract with the project for several months exploring how to use SQL Server Integration Services (SSIS) to accept records as XML documents over the Internet. Bitterlich was able to populate the administrative database using SSIS but faltered in attempts to create AIPs outside of BizTalk. In June 2010, Richard Pearce-Moses left the PeDALS project as principal investigator, taking a new position at Clayton State University

in Georgia. Watters, who had been the PeDALS project manager, assumed the principal investigator role. Slawomir Mitak, the PHP programmer in Florida who wrote most of the administrative catalog web interface, found a new job in fall 2010.

Aside from the personnel changes, 2010 was a time of great technological progress for PeDALS. Planning for the public catalog commenced in January 2010 and work continued through the year. A sub-grant from the PeDALS grant was given to South Carolina, in April, so Guzzi could continue working on developing the public catalog beyond his normal work and during his furlough weeks. Much of the spring of 2010 was devoted to ensuring that the original partner states had their BizTalk and SQL servers installed correctly and helping Alabama and New Mexico learn the methodology and begin the installation of their servers.

In February 2010, Microsoft released the technical specifications of PST files, which had been proprietary until that point. Schnackel spent the next six months developing a tool that would convert all messages in a PST file into individual XML records along with file attachments. He also worked on standardizing the BizTalk code that Neudesic had provided by the end of their contract in June 2010. Schnackel created a formal methodology of deploying BizTalk code for all partner states. This enabled code written in one state to be deployed easily on servers in another state and was the first significant step in creating reusable code for the PeDALS partners.

Although Schnackel had hoped to have the PST parser ready for the July NDIIPP partners meeting in Washington, D.C., he released it to Sourceforge in September 2010, in time for the Best Practices Exchange in Phoenix. At the October 2010 BPE, Schnackel and Matt Montano of New Mexico, demonstrated that the BizTalk schema mapping could be replaced by a much simpler process, with XML files generated through Access and Excel from index files provided with the records. Despite not having much BizTalk or .NET experience, Montano was able to create BizTalk applications to quickly ingest several record series using this method.

During 2010, Watters created a virtual development environment for the PeDALS technical teams, which allowed programmers in each state to work on BizTalk code outside of their production environment. Prior to this, most of the partners performed all development and testing on the production servers, requiring frequent rebuilds and re-ingesting of record series as programming and processes were improved.

While the technical work was being done, archivists, librarians, and records managers in the partner states were collecting record series to test in their PeDALS implementations. Sub-grants were awarded to South Carolina, New Mexico, and Wisconsin to pursue series that would be candidates for PeDALS ingest and to finesse metadata. These sub-grants greatly assisted states in continuing to work on PeDALS as the economy continued to cause states to reduce staff and funding to their departments.

South Carolina was particularly successful in finding a variety of records from different agencies, primarily because they had involved their records management staff early on in the process. New Mexico saw an opportunity to use PeDALS as a preservation tool for government publications, and identified the New Mexico Register as the first record series to be ingested. New York, which experienced severe budget cuts in 2009-10, had lessened its role to being an observer of the project earlier in the year. In September,

however, it identified a series of aerial photographs that would be ideal for a PeDALS private LOCKSS network, making it possible for New York to work more actively with the partners again. Wisconsin had also collected a number of record series from various state agencies and selected Department of Transportation Railroad files as their initial test case.

Before he left in June, Pearce-Moses submitted a grant request to NHPRC for continuation of PeDALS as a collaborative partnership beyond the life of the NDIIPP project. In September 2010, the NHPRC sent review questions to PeDALS, in response to the grant request submitted by Pearce-Moses. The primary concern was how PeDALS planned to sustain itself as a community of practice. Pearce-Moses and each of the partner leads responded to the questions individually, with the exception of Florida. A response was drafted and vetted by all the partners in a collaborative response. The NHPRC did not fund the PeDALS grant request. By practice NHPRC does not specify reasons for not funding particular grant applications. While attempts to receive grant funding for continuing PeDALS were not successful, project staff are dedicated to continuing the project (see Planning section).

Both Alabama and New Mexico struggled through the summer and into the fall of 2010 with installing and configuring their BizTalk, SQL, and admin servers. It took a site visit to New Mexico in September to ensure that the servers were set up correctly. By October, all states with the exception of New York had working production and development PeDALS environments.

PeDALS partners started to migrate their LOCKSS network servers from OpenBSD to CentOS in the fall of 2010, because LOCKSS was dropping support for the older operating system. Wisconsin and Arizona took the lead to update the LOCKSS network documentation. Arizona then worked with Tom Lipkis at LOCKSS to ensure the method and architecture would work for all partners. New Mexico and Alabama, however, had newer, more powerful LOCKSS servers based on the type used at MetaArchive, and these required BIOS and firmware patches to work. Cody Misplay in New Mexico worked with LOCKSS technicians to determine how to fix the problem. A site visit to Alabama in March 2011 by Watters ensured that their LOCKSS network was set up correctly.

2011

While Guzzi had finished writing the public catalog in ASP.NET in the Fall of 2010, a major sticking point was the lack of an advanced search capability. This was resolved by January 2011. All of the pieces to a PeDALS system were in place with the exception of an automated process to create dissemination information packages (DIP). Schnackel turned his attention to the creation of this automated process using the Windows Communication Foundation API to tie business rules in BizTalk with libraries that would create DIPs in various formats, most notably PDF. Up to this point, DIPs were created using the libraries via manual execution of a separate Windows-based user interface, not time-consuming, but hardly automated. He used the user interface from the OCR stripping program he developed for Alabama and plugged in the DIP processor to create a semi-automated process.

In March, 2011, a partners meeting in Arizona, allowed the partners a chance to review the project as well as receive training on the new templating system that Schnackel had developed in order to streamline the PeDALS ingest process and make it more ubiquitous across agencies. Preliminary discussions were also started regarding the

eventual transfer of LOCKSS servers to partner agencies in order to create more widely distributed Private LOCKKS networks.

In 2011, staff turnover continued to plague the project. South Carolina project lead Bill Henry retired in spring 2011, and Guzzi found a new job at another South Carolina agency shortly thereafter, leaving only Bryan Collars to work on PeDALS in that state. In New Mexico, Angela Lucero, the original project lead, left the agency just a few months after her state joined PeDALS. John Martinez assumed the role of project lead in New Mexico. Matt Montano, New Mexico's original technician was replaced by a new hire, Cody Misplay. In September 2011, Watters left the project as principal investigator. Linda Reib, the Electronic Records Archivist at Arizona State Archives, replaced Watters as the new principal investigator. This churn effectively slowed, but did not stop progress as new personnel were brought up to speed on workflows and the technical side of the project.

In late December NDIIPP granted the PeDALS project a two month extension in order to develop and test the ingest process of Arizona court records. Court records were selected as they are transferred to nearly all state archives throughout the county and territories.

2012

Schnackel completed work on the automated DIP processor by mid-February 2012. Work for deployment of this automated process and its required system modifications continued into March 2012. Much of the later work on the automated DIP processor occurred in parallel with the pilot record series project, with the Arizona Supreme Court, to determine compatibility of the PeDALS system for archiving court records exported from OnBase, a common document management system. The results of this pilot project proved a positive match for the process model and workflow of the PeDALS system, and no technical impediments were shown to exist during the ingest process of test data. However, due to the limited metadata provided by the courts much of the metatdata needed to be added manually. During further discussions with court technical staff, it was found that most metadata needed for effective description and access was held in a separate case management system. Work will continue with the courts in order to development tools and processes needed to pull documents and metadata from separate servers.

Overview of the system architecture

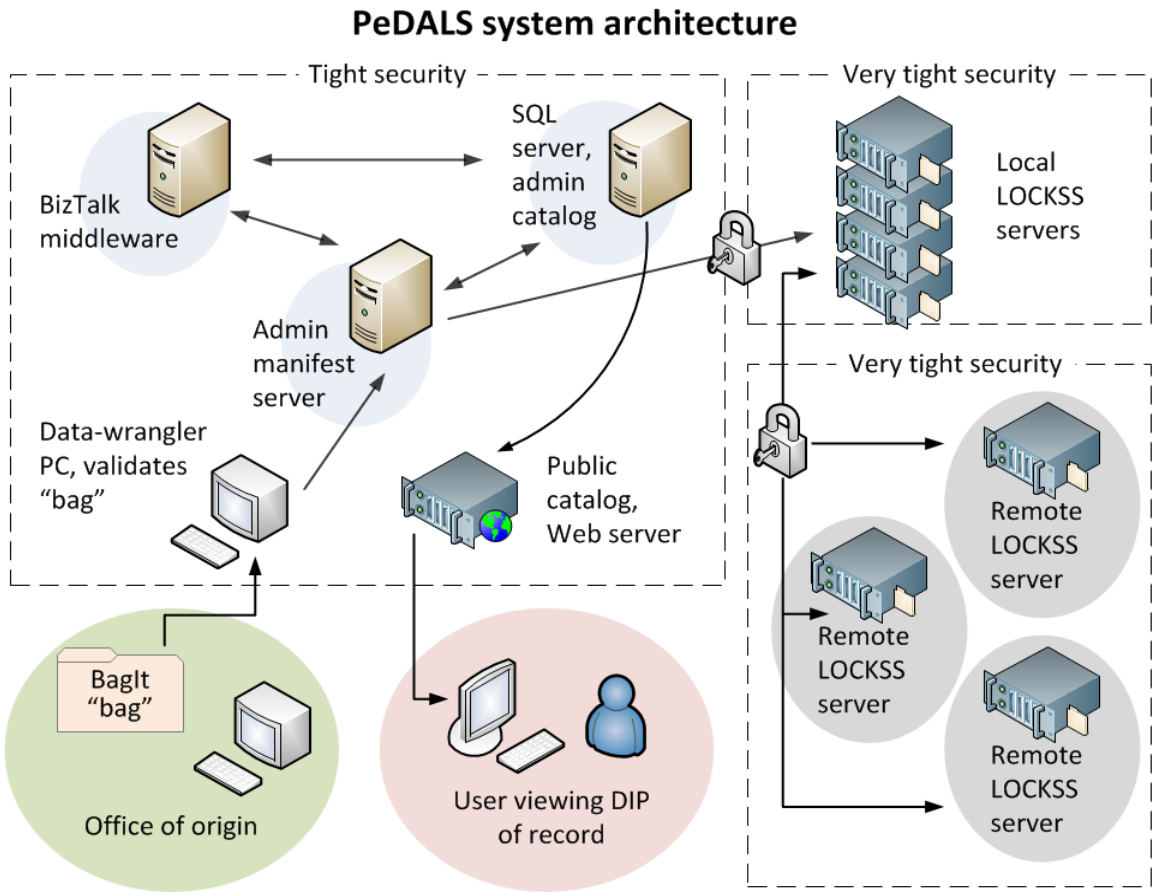
Using the framework suggested by the OAIS reference model, PeDALS is a modular system with different components to handle submission, archival, and dissemination information packages. Following a set of digital records through the PeDALS process can provide a quick overview of the function of each of its components.

Ideally, digital records and a metadata-rich index for the records are first introduced to the PeDALS system at a "drop box," a normal PC with anti-virus and normal text and spreadsheet editing software. In this scenario, the office of origination transferring the files will provide the digital records in BagIt "bags" (http://sourceforge.net/projects/loc-xferutils/), a data-transfer specification developed by the Library of Congress NDIIPP partners, which provide tools to validate the authenticity and completeness of the records being submitted. The files can be transferred via FTP, on optical media, or a hard drive. Once the records are in the drop box, an archivist can begin the process of "data wrangling" – validating the completeness of the files using BagIt, checking that the files are not

corrupted or infected with viruses, and looking at the metadata with an eye toward determining how it can best be used when writing the business rules used by the middleware. The records are also run through the New Zealand Metadata Extraction Tool (http://meta-extractor.sourceforge.net/) to glean additional preservation metadata.

After the records are vetted, the archivist enters high-level information about the records into an administrative catalog database running on a Microsoft SQL server. This information deals only with the aggregate – data about the provenance, the series, or the specific acquisition set – not the individual records. The archivist then discusses business rules about the records with a technician, for example, how best to enrich the metadata, what dates can be set for opening records, or how to handle dissemination copies of the records. The technician translates these business rules into code on a Microsoft BizTalk server.

Once the coding is complete, the records and accompanying metadata files are exposed to the BizTalk process, which applies the coded business rules, updating the administrative catalog database with item-level information on the records and creating an archival information package (AIP) that wraps each binary-encoded record with metadata conforming to the PeDALS metadata standard. These AIPs are collected into a "superpackage" that contains the metadata and records from the submitted set. The archivist checks the processed records using the same web tool used to enter high-level metadata and when satisfied with the accuracy and completeness of the package, accepts the submission. Only then does BizTalk finish updating the administrative database, populates another database that the public can search, and creates dissemination copies of the records for the public if applicable. BizTalk also exposes the superpackage to the PeDALS LOCKSS network, which harvests them for preservation.



The original PeDALS architecture included the servers noted in the diagram above, as well as a utility server that was used to create LOCKSS plugins and serve as the point where superpackages could be exposed to LOCKSS. The current architecture moved where the superpackages are exposed, from the utility server to the manifest server, thereby eliminating the need for a separate utility server solely for exposing superpackages for LOCKSS ingest. Some partner states have consolidated the administrative and BizTalk servers on the same physical machine. Also, rather than have duplicate servers for testing code and ingests, all states have development systems on a single virtual machine using Microsoft Virtual PC. This suggests that all but the PC used as a drop box and the LOCKSS networks could be virtualized for additional efficiency and savings in hardware and energy expenses.

Technical project findings

PeDALS will work for sets of homogenous records that have some form of index created by whatever standard applications the originating office used to administer the records for their business life. For example, an ideal set of records for PeDALS would be a collection of PDF files (or other singular format) that has an Excel spreadsheet or XML file listing the contents of the collection and corresponding descriptive metadata. The reality, however, is that the records archives find most interesting are rarely that structured. All partner states have received troves of unstructured digital data – hard drives from governors leaving office, for example – that are extremely valuable from a historical perspective, but do not fit the model that PeDALS prefers. To further complicate matters,

the introduction of one stray file into a record set can interrupt the PeDALS process. A hidden thumbnail-preview file (seen as "thumbs.db") in a BagIt bag of JPEG files, for example, would require that the bag be resubmitted by the originating office after the offending file was removed. This is not to suggest that PeDALS will not work; it does work if the records are carefully selected and vetted before they are ingested.

The initial PeDALS design planned to provide an inexpensive method of preserving electronic records. The initial costs for hardware and software is low when compared with off-the-shelf systems. Ongoing costs of hardware and software licenses, maintenance and server upgrades must be considered. The continuing expense of critical staff to manage the systems, both archivists who understand digital materials and manage digital workflows and technicians who can program in BizTalk and manage enterprise-grade servers was more expensive than anticipated. Government entities are currently facing budget cuts, making it difficult to hire and keep the skilled staff PeDALS requires. While PeDALS is not as simple a system to grasp or manage as was initially envisioned, participating states do have the staff and the basic structure in place for a working digital repository system.

A major highlight of the project was the collaborative spirit and consensus that arose out of creating a common workflow. Archivists from each of the partner states recognized that the tools that have served them well in handling historical documents and artifacts could be transferred to the technology required to process digital objects. When additions were deemed necessary by the archivists and librarians, changes were made to both the metadata dictionary and the administrative database. For the most part, these changes reflect the archivists' care in defining the workflow. Other changes were necessitated by the technology, such as defining in the database whether a record has had a dissemination package created.

PeDALS uses two open-source tools, BagIt and NZME, to authenticate and validate records and to provide additional metadata used by the BizTalk process in applying business rules. Of the two, the BagIt program has proved invaluable to archivists in ensuring the integrity of electronic data transferred to the custody of the PeDALS system and in providing an accurate count of files that are to be ingested. Even in the case of data "dumping", BagIt was used to create identifying hash values on files contained within non-indexed hard drives turned over to archives staff. Archivists in Alabama were pleased that nearly a terabyte of data could be transferred via BagIt to a hard drive, although they were not surprised that this took several hours to complete. The second open-source tool PeDALS depends on is the New Zealand Metadata Extractor (NZME), which provides preservation metadata about the records (information about the file format, size, creation software version, and creation dates). PeDALS had originally considered JHOVE (JSTOR/Harvard Object Validation Environment) for this purpose, but found that NZME recognized Microsoft Office formats more accurately than JHOVE for our program. NZME was easy to adapt when necessary. For example, when PDF files were discovered in South Carolina to have different types of date formats, it was simple enough to change the Java code NZME was written in to ignore the date differences, allowing the records to be ingested. The code solution was submitted to NZME and included in their source code in March 2011. Also, NZME relies on XML transformations that are simple to modify and update, and PeDALS has done so for some types of objects, notably GIF files in email attachments.

Findings related to BizTalk

Microsoft's BizTalk middleware is intended to manage transactions and communications between disparate systems by way of XML messages. It is ideally suited to handle thousands of small, asynchronous events and is used primarily for hotel reservations, factory inventory, and other business "just in time" systems. Although not necessarily intended for continuous batch processing, BizTalk can work in that manner, accepting messages from various locations, acting on these messages, and returning messages at another location. In a nutshell, that is what it does for PeDALS. It waits for messages that a record set is ready – the BagIt manifest file and New Zealand metadata XML file in most cases – and then cycles through each individual record listed in the manifest, constructing an AIP that normalizes all the metadata associated with the record, wrapping this metadata around a binary of the record, and appending each record's AIP into a superpackage encompassing all AIPs in the record set. At the same time, it populates the administrative catalog SQL database with the normalized metadata. Once finished, BizTalk waits for a message from the archivist that the set is well-formed, and only then releases the superpackage for harvesting by LOCKSS.

The BizTalk middleware code was originally written by consulting programmers from Neudesic (http://www.neudesic.com), a company that specializes in developing systems using Microsoft technology. The first successful BizTalk solution for PeDALS created AIPs for marriage certificates and used two applications: a PeDALS application common to all series and an application specific to the file type and metadata associated with marriage certificates. Three different programmers worked on the solution in turn. The third programmer also worked on the second solution for PeDALS, a series-specific application that would handle email parsed from an Outlook .pst file. Both of these BizTalk solutions worked for the problems presented – marriage certificates and email – but they did not further the project's larger goal of creating reusable code. The BizTalk email solution, for example, worked well on email, but it did not handle all XML records that contained self-referential metadata.

The primary challenge was to standardize the .NET environment for all partners using the same BizTalk implementation so that code written in South Carolina, for example, could be installed and run in Wisconsin. Each state had its own application directory, which ran counter to the principle of shared code. Schnackel overhauled the BizTalk environment, replacing each state's directory with a single directory for series-specific applications. Once finished, the partners could run each other's applications without issues, important if one were to adopt one state's application and make minor changes for a similar series.

Both the marriage certificate and email code were written on the assumption that each individual record in a series would have a corresponding XML file with descriptive metadata – a one-to-one relationship (in the case of email, the message was itself that XML file). More likely, though, were cases that a set of records would have a single file containing descriptive metadata. This was, indeed, what partners were finding when they solicited records. It soon became apparent that there were three basic types of applications that would handle most sets of records, as long as the records were uniform and not a mixture of file types: records with accompanying individual descriptive metadata files such as the initial set of marriage certificates (one-to-one), records with a single index file of descriptive metadata (many-to-one), and records that could be expressed as XML that included its metadata such as email and some database records (self-describing). Ultimately, this was refined further, as a utility to join individual metadata files into a single index file made the

one-to-one scenario unnecessary. Writing a basic application for both of the remaining cases created the reusable BizTalk code the project had hoped for and removed the need to write code for each series that presented itself. Modifying the two applications is much less troublesome than creating applications from scratch and still accommodates series-specific differences in business rules and metadata mapping.

Once these two core applications were defined, Schnackel designed a workflow that allowed schemas and orchestrations to be created outside of BizTalk using simple tools like Microsoft Access and Excel. The resultant XML files are easy to understand and import neatly into BizTalk. Using this technique, most typical errors are caught during the import process instead of in the middle of a BizTalk run, saving hours of time trying to troubleshoot what would invariably be an inadvertent typo. The workflow also leveled the playing field for the partners, regardless of their BizTalk expertise. Before the workflow, only Arizona and South Carolina had created their own series-specific applications. The other partners used the workflow to create applications rather than going through the trouble of managing BizTalk schemas and orchestrations.

PeDALS was successful in using BizTalk as middleware in its system architecture. Key findings while working on this project include:

- Relying on a consultant's expertise can be expensive, and deconstructing the code they wrote proved to be time-consuming but ultimately necessary. High turn-over rate of assigned programmers can slow the project and unnecessarily complicate the code.

- The current umbrella BizTalk application deals with the superpackage as its final message. This constraint is a limitation of LOCKSS/UNIX regarding cluster file counts (see paragraph 3 of *Findings regarding LOCKSS*) and was coded into the application by the consultants. It would have been preferable for BizTalk to send individual AIPs to a holding directory and have a separate process append them into a superpackage once the record set is approved. The amount of processing time BizTalk takes to create a superpackage with more than one thousand records can run several hours. There is a small cottage industry devoted to optimizing BizTalk applications to improve processing speed and its best practices are currently being investigated.

- There is an effective size limit for an individual record of about 750MB set by the amount of memory that can be addressed by the 32-bit Microsoft .NET framework executable environment used by the BizTalk process for the Base64 binary-to-text transformation necessary for storing binary files in XML. Moving to a 64-bit executable environment for BizTalk removes the limit of addressing space inherent in 32-bit systems and makes the file size limit dependent on the amount of RAM and virtual memory available to the BizTalk server instead. For example, a 64-bit BizTalk server with 8GB of RAM and the same amount of virtual memory would be able to process a 5GB record, given the amount of overhead needed to hold the record and its metadata while creating a binary of the record for encapsulation into the AIP.

- Some processes are more easily handled outside of BizTalk than within. The first iteration of BizTalk code was going to create a hash value for each record, but this is much easier to accomplish with BagIt and more beneficial to ensuring the trustworthiness of the record. Similarly, the process of creating DIPs can be launched by BizTalk based on business rules, but the program that accomplishes this is independent

of BizTalk and relies on code written by the PeDALS team and a third-party .NET library that creates and manipulates PDF files.

- BizTalk is complex and difficult to learn. Microsoft recommends two courses, one five-day course for programming and one two-day course for configuring and maintaining the system. The expectations of those taking the course are that they have sufficient .NET and MS SQL experience (http://www.microsoft.com/learning/en/us/course.aspx?ID=2933A&locale=en-us#tab3), not common skill sets in libraries and archives, especially in IT shops that are used to open-source software and systems. Watters and Schnackel found that working directly with partner developers using GoTo meeting was the simplest way to reinforce the BizTalk skills that otherwise could have gone fallow after training. Twice-weekly technical team calls and ad hoc remote problem-solving sessions allowed those who participated to see how BizTalk should be implemented.

- BizTalk and other enterprise-grade software from Microsoft is expensive. Pearce-Moses negotiated with Microsoft to get academic select software licenses for the PeDALS partners, which greatly reduced the cost of BizTalk and other server software. PeDALS paid approximately $3,900 for BizTalk and other licenses that would have cost nearly $24,000 for non-academic entities. A single-server license for BizTalk that allows five applications to be hosted is $10,138. PeDALS purchased the same license for $1,580. This cost may make it prohibitive for entities that do not receive academic discounts, although most libraries can qualify for the discount.

Findings regarding LOCKSS

All tests show that LOCKSS can be used to store AIPs in a secure, distributed fashion that diminishes the potential for bit rot or file corruption. Although it does not solve all of the problems presented by digital preservation (file type migration, for example, is not a LOCKSS feature), in the PeDALS system, it does preserve the bitstream of digital objects and relevant metadata for future solutions and provides access to the original record. PeDALS employs seven servers in a cluster, per LOCKSS's recommendation, for each partner's private LOCKSS network (PLN). The PeDALS model has geographic distribution of some of these servers to other partner states, to lessen the potential damage from local catastrophic events.

Another benefit of LOCKSS is that the hardware costs for its servers are notably less than the costs of enterprise-grade storage systems. Because LOCKSS is intended for storage instead of retrieval, it can rely on slower, less-expensive drives. Having seven copies of objects means that if a disk fails, it can be swapped out with a replacement (currently a 2TB replacement drive costs less than $100) and repopulated through polling over the course of a couple days, even if the server is on the other side of the continent. In addition, recent upgrades in server capabilities from Iron Systems from 1 or 2 TB drives to 16 or 24 TB systems has greatly enhanced the ability of archives to take in larger record accessions.

The original intent for PeDALS was to submit individual AIPs into the LOCKSS servers for each digital object, but Unix/Linux has a limit to the number of addressable files it can efficiently manage. LOCKSS advisors were concerned about the processor overhead of polling multitudes of small items instead of aggregates of these items. It was decided that bundling all of the AIPs from a single ingest into a "superpackage" would be easier for

LOCKSS to handle. LOCKSS checks the hash on the superpackage, which will show a discrepancy if any portion of any of the AIPs within it change, not just the bitstream of the digital object, but the metadata as well. PeDALS can access individual AIPs through the superpackage URL easily and restore the object as necessary, checking the hash for the object in the administrative database, or the hash in the AIP itself.

The challenges we encountered with LOCKSS tended not to be technical in nature. Once a PLN is set up, it requires much less maintenance than more expensive enterprise storage systems. LOCKSS is capable of ingesting large superpackages and polling servers across the nation. A South Carolina LOCKSS box located in Arizona, for example, would show a completed 1GB superpackage within five hours of its being exposed by PeDALS in South Carolina. We tested superpackages totaling nearly 1.5 TB for polling across a network without issue. The challenges we encountered with LOCKSS include:

- Minimal documentation from LOCKSS on specific functionalities. Although there is quite a bit of information on setting up servers as part of the global LOCKSS network, there is limited information on how to establish a PLN. The documentation PeDALS has on configuring its PLNs is the culmination of trial and error, advice from LOCKSS technicians and architects, and research on similar, self-documented PLNs, such as those of the MetaArchive. The PeDALS documentation for establishing a PLN has been sent to LOCKSS for vetting. We have installed five PLNs using our documentation.

- Some of the partner states express the concern that LOCKSS may be inefficient and potentially expensive method of storing the rapidly expansive amounts of government data. The original PeDALS partners experimented with 2 terabyte or smaller LOCKSS clusters. Alabama and New Mexico introduced 16TB LOCKSS clusters into their PeDALS systems. The rule-of-thumb data storage equation for an individual record is approximately 3.0x original record byte count + 2x the descriptive metadata byte count. We are presuming that we theoretically use all the descriptive metadata provided to the system, in addition to storing the provided original metadata in its own XML tag. Although the amount of data ingested currently is far less than these limits, Florida requested a cost analysis for 64 terabytes of storage, which they estimated they would need for audio data only, over two years. The resulting expense seemed prohibitive – not so much the initial cost of the 28 servers necessary at the time to store 64 terabytes, but the costs associated with leasing rack space for multiple servers from the state mandated server centers and paying maintenance on them.

- While understandably necessary, the annual LOCKSS Alliance membership cost that provides access to technical support can be problematic for some of the partners' budgets. The alliance fees normally are based on the type and size of a university that would be part of the global network (see http://lockss.org/locksswiki/files/LOCKSS_ALLIANCE_MEMBERSHIP_FORM.pdf). PeDALS negotiated with LOCKSS to set membership fees at $5,000 a year, about the rate expected of a large, non-research college. PeDALS partners were told in the fall of 2010 that they would be expected to pay their own alliance fees and to budget accordingly. Some of the partner states, faced with severe budget reductions and staff cuts, decided to forego the alliance membership, trusting they could maintain their individual PLNs without assistance. Other partner states remain in the alliance.

- Geographical dispersion of LOCKSS servers was a problem as some state agencies were hesitant or statutorily unable to allow records to be located in another state, regardless of the security measures put into place. Although the original participants in PeDALS agreed to distribute servers amongst themselves, once the LOCKSS servers were working and accepting records, changes in some of the states brought the practice into question. New Mexico found that it would take legislation to allow such distribution. New Mexico and other states decided to pursue the geographic distribution within the state as an easier option.

- Partners were also concerned that there may be imbalances in costs associated with interstate distribution, particularly as states consolidate IT infrastructures among multiple agencies to save costs and agencies are charged on a per-server basis. A state that may have only one server to distribute might not be able cover the costs to host three servers from a larger state. To address these concerns an agreement drafted by the partners dictates that that any such additional costs would be borne by the state owning the server. Partners that can distribute servers have indicated that they will abide by this agreement.

Key gains by partner states

PeDALS offers unique and necessary tools for digital records preservation. PeDALS meets a critical need for partner states in the PeDALS network. Partner states gained invaluable knowledge and abilities in the preservation of historic state government records.

Alabama

The value of PeDALS for ADAH was that it allowed the Records Management staff to focus on electronic records. The staff liked that the PeDALS Project was a practical attempt, not a theoretical construct, to deal with electronic records. The relatively low cost of the PeDALS software and hardware, when compared to other available systems, was also a draw. Had Alabama not participated in PeDALS, staff would not have acquired the needed skills necessary to address the preservation of electronic records and feel comfortable enough to use those new skills with state and local agency personnel.

The records Alabama staff first worked with on the project were photos and documents from the Commissioner of the Department of Agriculture and Industries. This group of records gave staff specific material to use in training state agency staff to transmit electronic records to the ADAH. Project staff created leaflets and training modules for state records liaisons on proper file naming techniques and acceptable file types.

The work with the PeDALS project has confirmed the ADAH staff belief in the viability of the LOCKSS process for continued dark archiving of master electronic files. Now included in the process of electronic records transmittals to the ADAH is BagIt - used to verify that records are transferred successfully and to authenticate them. Staff also created an electronic records transmittal form and a hardware receipt form (used when agencies transfer records to ADAH on their portable hardware devices).

Finally, one area not often highlighted in grant reports is the value received from attending national conferences, such as Best Practices Exchange (BPE) and NDIIPP. These provided the ADAH staff with valuable insight into the issues involved with electronic records preservation and in creating an e-records program. Staff came to understand just

how truly complicated, time consuming, and expensive the process of preserving electronic records is and can be. They are now committed to programs that emphasize collaboration and sharing.

Arizona

The Arizona State Library, Archives and Public Records (ASLAPR) staff dramatically increased their technical skills and ability to preserve electronic historic records of the state. Agency staff gained considerable knowledge as to the need for a digital archive, the technical complexity of that system, a realistic view of the cost, as well as developing a viable vision and roadmap for how to move forward from both a technical and business perspective. Considerable technical knowledge has been gained in the ability to relate traditional archival skill sets to the electronic record by looking for data field patterns, discovering how to structure programming rules based on these patterns, and then creating the program map to allow the computer system to process the records automatically with limited archivist intervention.

Through the PeDALS grant, staff was able to construct a basic electronic records preservations system and ingest several records series, to include the Athletic Training Board Director's email, county marriage certificates and a Supreme Court case. Working with agencies in order to understand the complexity of a record series stored electronically, including the metadata, lexicon, formats, life cycle, relational database structure and multi-server systems was more intensive, complex, and multifaceted than originally anticipated. We learned how in-depth the need to train record creators in archival concepts in relation to electronic records and PeDALS programs goals really was. This understanding will greatly enhance our ability to plan and move forward with our electronic records preservation program.

South Carolina

The SCDAH gained invaluable knowledge in the requirements necessary to properly ingest, maintain, and make accessible electronic records. An unexpected benefit was the experience gained in negotiating with State Agencies on the transfer of their electronic records to the Archives for permanent retention.

The SCDAH ingested a variety of series the biggest being the Circulars and Orders of the Public Service Commission, comprised of 4656 individuals files in PDF format with total size of just over 4 GBs. An additional 15 series were also ingested, comprised of 823 individual PDF files with a total size of 86.42 MBs.

During the latter stages of the project the SCDAH was adversely affected by decreasing budget issues that forced us to curtail our involvement and ultimately resulted in the project being put on hiatus.

New Mexico

Participating in the PeDALS project allowed the State Records Center staff the opportunity to gain knowledge in the following areas:

- Metadata definition, extraction, and assignment to records.
- Long term storage of records, using the LOCKSS servers, configuration of these servers.
- Record Ingest process- identifying record series, metadata compilation and extraction, preparation for ingest, ingest records and error correction.

The State Records Center and Archives ingested five record series into PeDALS:

- NM Register (192 Files, PDF, 96.6 MB)
- Executive Orders (454 Files, PDF, 43.8 MB)
- Fish Stocking Reports (348 Files, PDF, 3.17 MB)
- Governor's Speeches (14 files, PDF and DOC, 1.03 MB)
- LFC Newsletters (93 Files, PDF, 41 MB)

New York

Owing to staffing issues, New York State could not participate fully in the PeDALS project. The State Archives staff most heavily involved in the project were periodically compelled to focus on sweeping freedom of information requests, the State Library suffered loss of staff due to reductions in workforce and attrition, and the staff member responsible for coordinating the installation of various PeDALS technical components was promoted to a senior administrative position and had to put aside most PeDALS related work.

In addition, the project evolved in unanticipated ways. The project's increasing focus on archival records as opposed to state government documents led the State Library to adjust its level of participation from active participant to observer, a role that the Library maintained until the project concluded in December 2011. Most significantly, New York State lacked access to IT staff with expertise in BizTalk, nor did it have the resources to cultivate that expertise; when PeDALS was first conceived, the partners anticipated that an experienced BizTalk consultant would be hired.

The New York State Library and the New York State Archives nonetheless learned a host of valuable lessons as a result of its participation in the project:

- Early on in the project, partners discovered that coordinating with agencies to obtain records series or document sets was a time-consuming process and required more time and resources than initially expected. Also, getting agencies to provide automated access to their databases was problematic in many cases.

- Partners also found that record sets and documents were often contained within legacy and proprietary document management systems at the host agency. Since each of these systems had their own unique set of issues, importing this data was challenging.

- Partners also found that record series and document sets that were available to use were mostly scanned as opposed to born-digital material. This was the opposite of what was originally anticipated.

- Building a digital preservation solution from the ground up requires an immense investment of time and effort and a willingness to improvise, troubleshoot problems, and adjust as circumstances change. All of the partners expected to devote substantial resources to this project, but none of them knew at the outset just how immense the commitment would be.

- The State Library found that development of the metadata schema for this project provided a clearer understanding of the commonalities and differences between electronic records and documents and how those shared/different traits will

influence electronic systems that will be developed to preserve and access these materials.

- The State Library also discovered that the scale of the project needs to be considered when deciding on participation. PeDALS was originally designed to ingest large amounts of electronic records from agencies that had records already contained in electronic systems. This presented a problem when dealing with electronic state government documents. The majority of our electronic state government documents do not exist in content management systems and the State Library does not have set schedules for acquiring them. Also, the amount of processing and business rule writing that needs to be done is only worthwhile when dealing with large data sets, not individual documents or small runs of documents.

New York found the experience of participating in the development of the metadata schema and witnessing the development of the project's technical infrastructure extremely rewarding. We now have a much clearer sense of what building an in-house digital preservation system entails and the staffing and technical hurdles that we must surmount, and are drawing from this experience as we evaluate other preservation and access options.

Wisconsin

PeDALS had a direct and very positive effect on the operation and capability of the Library-Archives of the Wisconsin Historical Society in that it provided a focus and a vehicle for training, collaboration and innovation in electronic records. When we first decided to participate in the PeDALS project, electronic records were just starting to appear in collections from Wisconsin state agencies. We recognized at the time that this would be a growing and long-term challenge for us as an agency. We also realized that the finances and staff resources in developing a system to manage the crush of electronic records was out of reach for us as an organization as it would be for most similar institutions.

The basic concepts behind Richard Pearce-Moses' initial proposal not only offered us the ability to work through the translation of archival concepts to the electronic record world with others facing the same challenges, but it also allowed us to take advantage of being able to share resources and skills between the states in a way that we would not have been able to do otherwise. Each state came to the table with specific knowledge and experiences that they were able to leverage into the project. Outside personnel were then hired to fill in the skill gaps that the collective lacked.

The PeDALS project has provided the Wisconsin Historical Society with a much improved understanding of the essential digital repository concepts and practices. Through working with our PeDALS partners as well as meeting with other collaborations at national conferences, we now have a firm understanding about what requirements are needed in building such a repository. We now have a well-documented system in place that is ready to be developed and enhanced in coming years, and we foresee PeDALS processes being the cornerstone of this system.

PeDALS Forward Plan

Going forward the PeDALS partnership will be a loose confederation of the remaining active states – Alabama, Arizona, and Wisconsin, with New York being an active observer. Due to continued funding and staffing issues, Florida and South Carolina will be

unable to continue participating. The New Mexico State Records Center and Archives will be pursuing other Commercial Off The Shelf (COTS) electronic content management software applications.

Arizona has committed to funding the PeDALS application (BizTalk) programmer through June 30, 2012. The Arizona Secretary of States' Office, which now over sees the State Library, Archives and Public Records, is pursuing a budget adjustment to fund the programmer for the 2012-2013 fiscal year. Wisconsin and Alabama have committed additional funds for the programmer during the 2012 calendar year. Alabama, Arizona, and Wisconsin are currently working on new memorandums of agreement in order for funding to transfer between states and to solidify our program goals going forward.

Monthly scheduled meetings have been set for administrative staff to continue working collaborations. Bi-weekly technical staff meetings are ensuring the continued development of PeDALS applications.

Moving forward by continuing states

Alabama

ADAH will continue to work on ingesting records series into PeDALS. Our first goal is to get preservation and access copies of the Department of Agriculture and Industries photos. The total size of the 2,000 Agriculture and Industries photos is 2 GB. Following the completion of that ingest, we will ingest other records series into PeDALS, such as the Governor's Executive Orders. We want to make the process of ingestion more automated.

We will collaborate with the PeDALS partners still interested in and capable of participating in PeDALS. We will contribute money to pay for Brian Schnackle in Arizona. In addition, we will collaborate with non-PeDALS partners, such as North Carolina State Archives. ADAH staff is convinced that only through mutual collaborations among institutions, states, and other entities will standards and practices be advanced and made practical for all.

We will continue developing our electronic records program. We plan on providing electronic records management training opportunities to state and local government agencies. While the Archives only takes in permanent electronic records from state agencies and no electronic records from local ones, we have an obligation to help agencies at all levels to properly manage their electronic records. One of the areas we need to address with state agencies is the inclusion of metadata with their records that allows for records to be ingested, accessed, described, and preserved.

Arizona

The Arizona State Library, Archives and Public Records is dedicated to continuing the PeDALS collaboration as well advancing electronic records preservation and public access to these records within the state. Our agency has formed an electronic records preservation systems working group to move PeDALS from the research and development stage into a full production system. This group is working on enhancing system design and functions, in conjunction with the PeDALS on-going collaborative. This group is also working on establishing reliable and on-going state funding for the system and related staff.

Arizona will continue to work with the state court system to complete development of tools necessary to pull records from document management systems and metadata from related case management systems. Work will continue on developing transfer policy and procedures for electronic records, preservation training for state and local government entities, as well improving the capabilities to ingest and make accessable state documents and other record series.

New York

New York is committed to remaining an active project observer. State Archives staff continues to take part in project conference calls and are currently helping to test collaborative tools that will enable the project to continue moving forward. State Archives staff are also exploring the possibility of setting up a PeDALS LOCKSS cluster; the State Education Department recently set up a Linux environment for another initiative, and as a result it may be easier to secure departmental support for LOCKSS technology. Owing in part to its PeDALS project experiences, the State Archives is also exploring the possibility of outsourcing the preservation and storage of many of its electronic records. State Library staff continues to monitor the PeDALS project via online collaborative project tools and participate in project conference calls when available. The State Library remains interested in the outcomes and lessons learned regarding LOCKSS.

The New York Archives is exploring the possibility of implementing a modified version of PeDALS using the Archivematica open-source project to create AIPs, which are ingested into a LOCKSS network, and possibly the PeDALS DIP processor to extract the records from LOCKSS. However, Archivematica is still in beta testing and staff has encountered numerous problems which will need to be address.

Wisconsin

The initial phase of the PeDALS project has succeeded in creating a streamlined system that allows the bulk ingest, trustworthy archival storage and presentation of digital content that follows the OAIS model in a way that simply digitizing and storing paper documents or storing already born-digital materials and then presenting the materials cannot. The Wisconsin Historical Society plans to continue utilizing the PeDALS systems as a cornerstone of its overall digital repository planning efforts that are currently under way. Wisconsin will also cooperate with Arizona and other states to ensure that the skills and software tools developed by the PeDALS effort continue to be effective and useful.

Currently, we are working on using our training and experiences with PeDALS to better manage the acquisition, and eventual ingest of large amounts of unstructured records from Wisconsin state agencies. We are currently experimenting with a large records set from a former governor to explore how to automate the curation and organization of the records so that they can be ingested into PeDALS later in 2012. As we develop our methodology and toolset, we plan on releasing our findings to our partners and then to the larger archival community.