

Sample Application

2008 National Leadership Grants for
Libraries

Demonstration

University of Florida

Towards Interoperable Preservation
Repositories (TIPR): A Demonstration
Project

TOWARDS INTEROPERABLE PRESERVATION REPOSITORIES (TIPR): A DEMONSTRATION PROJECT

ABSTRACT

The Florida Center for Library Automation at the University of Florida, and its partners the Cornell University Library and the New York University Libraries, propose a two year demonstration project titled "Towards Interoperable Preservation Repositories (TIPR): A Demonstration Project."

The task of preserving our digital heritage for future generations far exceeds the capacity of any government or institution. Responsibility must be distributed across a number of stewardship organizations running heterogeneous and geographically dispersed digital preservation repositories. For reasons of redundancy, succession planning, and software migration, these repositories must be able to exchange copies of archived information packages with each other. Practical repository-to-repository transfer will require agreed-upon transfer protocols, enhancements to repository software applications, and above all, a common, standards-based transfer format capable of transporting rich preservation metadata as well as digital objects.

The three university partners address these needs in this demonstration project. Each partner runs a digital preservation repository, but each of the three systems are based on different software platforms and specialize in different types of content. Building on prior work, this two-year project will define a common transfer format, modify the three repository systems to import and export information packages in this transfer format, and test a carefully developed set of use cases to verify the usability and flexibility of the format.

This project will provide a proof of concept for the exchange of information packages, a model for transfer, a standards-based transfer format, and information about issues likely to be encountered when transferring information packages from one repository to another. The transfer format will be based on standards currently used by the preservation community, including PREMIS (for preservation metadata) and METS (for package description). Guidelines for using PREMIS and METS in this context will be forwarded as formal recommendations to the PREMIS Editorial Committee and the METS Editorial Board. An XML schema definition and documentation for the transfer format will be published and made freely available to the international preservation community, as will the formal project report.

This demonstration project will not be the last word in repository interoperability, but will advance the state-of-the art significantly and provide a new baseline for future work to build on. As a result, future preservation repositories will be demonstrably more interoperable and better positioned to be certified as trustworthy.

TOWARDS INTEROPERABLE PRESERVATION REPOSITORIES (TIPR): A DEMONSTRATION PROJECT

1. Assessment of Need

If there is one thing that everyone involved with digital preservation agrees on, it is that responsibility for digital preservation can not be centralized, but rather must be distributed across a number of heterogeneous, geographically dispersed, stewardship organizations. Factors arguing for a distributed approach include the massive volume of at-risk materials, the technical differences between formats and media, the range of functional needs in different communities, and the applicability of different political and legal regimes. Also, there is a strong belief within the preservation community that there is no single "true" preservation solution, that many approaches must be tried and tested, and that redundancy reduces risk.

If no one institution can preserve everything, however, neither can wholly isolated and independent organizations. Nationally and internationally, there is a focus on mechanisms for cooperation, coordination, and federation of preservation efforts, as well as the development of shared standards. In the United States, the National Digital Information Infrastructure and Preservation Program (NDIIPP) program is taking the lead in establishing a distributed digital preservation network. [1] In Europe, the PLANETS (Preservation and Long-term Access through Networked Services) project funded by the European Union is a major vehicle for integrating distributed preservation services.[2]

Cooperation is needed on many levels, from the organizational to the technical, and within many domains, from archives to institutional repositories. One concrete need within all domains is for preservation repositories to be able to exchange stored information packages with each other. (An "information package" is a unit of preservation including both content and metadata. Various types of information packages are defined by the Open Archival Information System reference model, a core standard within the preservation community.[3] This ability is needed for several practical reasons.

First, no repository is risk free. Any given repository may employ a flawed preservation strategy (e.g. a lossy migration) or may lose data through disaster or negligence. Optimally, digital content of high value should be preserved in more than one repository to increase the chances of survival over the long term. As there are few preservation repositories in production at this time, it is likely that content initially stored in one repository will want to be copied to other repositories as they are established in the future.

Second, trustworthy repositories are required to have a succession plan in the event they cease operation for any reason. [4] In many cases the preferred plan would be to transfer content from the original repository to another, more successful repository.

Third, preservation repositories rely on software applications, and no application lasts forever. Just as with integrated library systems, custodial institutions will want to change repository systems over time to take advantage of better features, lower costs, or newer technologies. Information packages stored in the old system will have to be migrated into the new one without loss of important preservation information.

In 2002, the Library of Congress and the National Science Foundation co-hosted a workshop to develop a research agenda for digital preservation. The identified priorities included research to "allow heterogeneous distributed repositories to exchange content and services." [5] This was addressed by one of the first NDIIPP funded projects, the Archive Ingest and Handling Test (AIHT). AIHT tested the feasibility of transferring a complete digital collection from one repository to another. The test collection consisted of about 57,000 files about the September 11th attack collected by George Mason University. In Phase I of the project, each of four participating institutions received the collection on a hard drive and ingested it into their own local repository. In Phase II, each institution exported its own collection and ingested the exported collection of a selected partner institution. The experiment, which was documented in detail in project reports and summarized in a special issue of *D-Lib Magazine*, provided enormously valuable information to the preservation community. [6]

The AIHT demonstrated that wholesale transfer of a collection was possible while exposing a number of conceptual and technical issues impeding portability. In this test, however, only the content itself was exchanged, with no metadata or processing information, and there was no attempt to create or use a common transfer format. In fact, the project chair concluded that one main outcome of the project was the need for further testing. [7]

A true preservation repository will require a certain amount of descriptive metadata to accompany content, and it will enrich ingested content by the addition of metadata of many types, including technical format information, rights and permissions, and processing history. The repository may create derivatives such as normalized versions to be stored alongside or in preference to the original object. It may create a version in a newer, more viable format. It must create metadata to record the events affecting the source object and its derivatives, and the relationships among the various objects. This family of objects and all of the metadata pertaining to it becomes the information package that must be preserved and transferred in usable form to the second repository, which will have its own ingest requirements, metadata schema, and preservation strategies.

The MathArc project advanced the work of AIHT by accounting for this more complicated environment. MathArc was a collaboration between Cornell University Library and Göttingen State and University Library that focused on the cooperative management of distributed digital preservation systems. In the design of this project, electronic journal content was to be stored redundantly in two partner archives using different repository software systems. A common data exchange format was developed using community standards, and an OAI-PMH based protocol was designed, but not

implemented, for automatically updating one archive when content is added to the other.[8]

While MathArc was ongoing, the German National Library devised a Universal Object Format specifically for exchange of archived objects between geographically and administratively distributed repositories as part of project kopal (Co-operative Development of a Long-Term Digital Information Archive). The project provides interesting information about general exchange issues, but is not immediately applicable because the repositories all use the same software platform. Also, Germany uses a different standard for preservation metadata than the English-speaking world.

Informed by AIHT, MathArc and kopal, Towards Interoperable Preservation Repositories (TIPR) will test the actual transfer of information packages between three technically heterogeneous, geographically distributed repositories: the Florida Digital Archive, based on DAITSS; Cornell University Library's CUL-OAIS, based on aDORe; and New York University Libraries' Preservation Repository, based on DSpace. The information packages will have been fully processed and enriched by the originating repositories and will be transferred in enriched form.

A key deliverable of the project will be the development of a common, XML-based transfer schema that builds on existing standards, including the Metadata Exchange and Transfer Standard (METS), PREMIS preservation metadata, and standards for format-specific technical metadata such as Z39.87 (MIX). METS is an XML schema that describes the structure of digital objects and has placeholders for inserting descriptive and administrative metadata. PREMIS is a data dictionary for core preservation metadata that has become a *de facto* standard within the English-speaking preservation community. There are many different ways that PREMIS information can be carried in METS, and a working group convened by the Library of Congress is expected to recommend guidelines in 2008. This project will test, amend and extend these guidelines as necessary, resulting in a usable community standard.

2. National Impact and Intended Results

This project will advance the state of repository-to-repository transfer from the relatively simple content employed in the AIHT to more realistic, complex and enriched information packages actually produced and stored by current preservation repositories.

For the first time, a common standards-based exchange schema will be developed and tested between distributed, technically heterogeneous archiving systems. This will offer the same type of advantage that the Z39.50 metalanguage offered for search and retrieval of distributed, heterogeneous catalog systems. With an accepted exchange schema, each repository in the international preservation network will need only develop export and import methods for a single foreign schema, rather than custom methods for each and every potential exchange partner.

Because the schema will be based on PREMIS and METS, issues involving the implementation of each of these standards for this use will be exposed. Results and recommendations will be returned to the PREMIS Editorial Committee and the METS Editorial Board, in order to feed into the maintenance activities for these standards. A standard way to represent PREMIS metadata within METS, and the proof-of-concept that PREMIS can be used to convey rich preservation information from one repository to another, will be extremely significant to the future adoption and use of the PREMIS specification.

The project will further the progress towards future certification of trustworthy digital repositories, as it will help to make it possible for preservation repositories to design succession plans that involve transfer of materials to other trusted digital repositories as required by *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)*.

Results should be applicable across a spectrum of interests, from the curation of science and social science research data to the archiving of commercial scholarly journals. However, it will be particularly relevant to libraries and archives, because these sectors has been actively leading preservation efforts in the academic and cultural heritage communities. Many of these institutions are participating in the implementation of preservation repositories that aspire to future certification as trustworthy digital repositories.

3. Project Design and Evaluation Plan

This project will explore the exchange of standards-based Information Packages between three digital preservation repositories. The primary goals are:

- to demonstrate the feasibility of repository-to-repository transfer of rich archival information packages as a strategy for redundancy, software migration and repository succession;
- to advance the state of the art by identifying and resolving issues that impede such transfers;
- to develop a usable, standards-based transfer format, building on prior work by Cornell, Göttingen, and the kopal project;
- to disseminate these results to the international preservation community and the relevant standards activities.

The project is designed to be carried out in four overlapping phases identified by their dominant activities: analysis, coding, testing and assessment, and evaluation and dissemination. (In practice, of course, these activities will be iterative. For example, an assessment of test results may lead to the identification of problems that require more analysis and subsequent recoding.) Based on the experience of earlier projects with geographically distributed partners, the schedule allows for the possibility that collaborative activities may take longer than planned, so ample time is allowed at the end for high-level assessment, evaluation and dissemination activities.

Phase 1, Analysis

Project partners will agree upon requirements for a transfer format, specify a common transfer format that meets these requirements, and agree upon one or more transfer protocols to be used. As part of the specification process, partners will develop use cases to be tested later on. Use cases will designate content with particular characteristics to be tested (for example, compound objects with and without hierarchy; objects with particular types of metadata, objects with multiple versions), how these should be represented in the transfer format, and expected results after broadcast and/or ring transfer (see Phase 3). Partners will map their local stored metadata to the common format and identify needed changes to their local metadata schema. Each repository will decide how best to produce an exchange package from its own stored data, and draft specifications for any development required. This phase will overlap with coding as approaches are prototyped.

Phase 2, Coding

Each of the project partners will implement the ability to export an archival information package in the common transfer format from their repository system according to specifications developed in Phase I. Each partner will implement the ability to ingest an information package in the common transfer format into their repository system. Both export and ingest must be designed to be as lossless as possible. For example, if the information package contains detailed provenance information provided by repository A, repository B must be able to retain that information and map it at the time of ingest to the format and values used internally for its own provenance information. Each partner will implement the ability to support the protocol(s) specified during the analysis phase.

Phase 3, Testing and assessment

Project staff will conduct a series of package transfers designed to test the use cases developed in Phase I. Two types of transfers will be employed. In a broadcast transfer, each repository produces one output exchange package which is delivered to the other two repositories for ingest. In a ring transfer, the output of repository 1 is ingested by repository 2, then disseminated from repository 2 for ingest by repository 3, then disseminated from repository 3 for ingest by repository 1. Results will be reviewed in light of expectations. Anomalies may lead to revisions in the repository systems, a revision of expectations, or analysis and documentation of the issues.

Phase 4, Evaluation and dissemination

These activities are described in more detail in separate sections below, "Evaluation" and "Dissemination".

Evaluation

A project of this nature can not be evaluated immediately in terms of its impact on its designated audience, the national (and international) digital preservation community. As stated on the IMLS website:

IMLS supports basic research, organizational enhancements, and other activities intended to strengthen the ability of organizations to provide high-quality services. Such projects may be designed to extend a discipline's knowledge or to create tools to improve practice, rather than to produce immediately observable benefits for end users. IMLS supports such projects because it anticipates that they will contribute to making lives better *in the long term*. In reporting results of such grants, IMLS wants to know what you believe long-term benefits will be for library or museum users and their communities, and how those improvements will be recognized when they're achieved.[9]

In this case the high-quality service ultimately strengthened will be the capacity of digital preservation repositories provided by or used by cultural heritage institutions. This project will provide a proof of concept for the exchange of information packages, a model for transfer, a standards-based transfer format, and information about issues likely to be encountered when transferring information packages from one repository to another. If these results have good take-up within the preservation community, future preservation repositories will be demonstrably more interoperable and better positioned to be certified as trustworthy. At a date roughly three years from the conclusion of this project, one would expect to see:

- the transfer format described by this project serving as a standard (formal or *de facto*) for the repository-to-repository exchange of information packages;
- any recommended changes to PREMIS and or METS incorporated into the published versions of those standards;
- a majority of the most commonly-used repository systems developing or planning to develop transfer functionality based on this model;
- a small but growing number of operational repositories with succession plans involving transfer of content to other repositories; and
- at least one future project funded to build on the work of this project, as this has built on earlier work.

Of course, short-term evaluation that does not involve an assessment of impact on the community will be carried out before the end of the project period. This evaluation will focus on whether the promised deliverables were successfully achieved. These include:

- testing of all use cases defined for the project;
- completion of all dissemination steps itemized below;
- schema definition and documentation for the transfer format, possibly as a registered METS profile;
- formal recommendations to the PREMIS Editorial Committee and the METS Editorial Board;
- a complete, formal project report, describing goals, methodology and results;
- documentation of issues impeding interoperability, and recommendations (where possible) on how these can be addressed.

4. Project resources: Budget, Personnel, and Management

The three partner institutions will each participate in the same activities: analysis and specification, modifying and enhancing their local preservation repository applications, conducting and evaluating the exchange of information packages, and summarizing and disseminating results. All three institutions are extremely well-qualified to undertake this project. All three have been active in the area of digital preservation for at least half a decade, operate working preservation repositories, are well known in the digital preservation arena. Each of the partners has received at least one prior grant for work in digital preservation, and each shares a serious commitment to advancing the state-of-the-art in preservation practice.

FCLA will use project funds to hire a programmer for one year, calendar 2009. Analysis and specification will be done by permanent staff who already understand PREMIS, METS, the generation of dissemination information packages, and the DAITSS repository application. This will be provided by FCLA as cost share. Franco Lazzarino (25%), a senior developer for DAITSS, and Randy Fischer (10%), the team leader for DAITSS, will provide the technical expertise. Together, they have 29 years of information technology experience. Priscilla Caplan (25%), Assistant Director for Digital Library services, will provide expertise in metadata, standards development and preservation community requirements. In addition, she will manage the project and coordinate the work of the three partner organizations. The University of Florida's cost-share will be over 100%, including two-thirds of indirect costs.

The Cornell University Library (CUL) will use project funds to support their Digital Preservation Programmer/Analyst Specialist, William Kehoe (55%), who will be responsible for both analysis and programming. Mr. Kehoe is already familiar with the Library's aDORe-based CUL Digital Archive, and with METS, PREMIS and the MathArc projects. All other contributions will be cost-shared. Mr. Kehoe will be supported by programmer/analyst George Kozak (5%), and metadata specialist Enrico Silterra (9%). All of CUL's participation will be under the direction of Oya Rieger, Interim Associate University Librarian for Digital Library and Information Technology. Ms. Rieger has extensive experience with technical metadata standards development and digital preservation.

Like Cornell, the New York University Libraries will use project funds to support an experienced staff member, Unni Pillai, for one year as their programmer on the project. All other contributions will be cost-shared. Additional analysis and programming support will be provided by Joseph Pawletko (10%) and Ekaterina Pechekhonova (10%), who have been deeply involved with the development of NYU's DSpace-based repository and have presented at Open Repositories, the Digital Library Federation Fall Forum, and several other conferences. Day-to-day oversight will be provided by James Bullen (5%), the repository Project Director. Bullen and Pawletko are also both 10% on an NDIIPP grant. The NYU PI for this project is Michael Stoller, the Director for Collections and Research Services. Dr. Stoller has been PI on grants from IMLS, NEH and Mellon. NYU cost-share for this grant will be over 100%, including contributions of the project staff and 90% of indirect costs.

5. Dissemination

Given the importance of this project and its goal of contributing to the development of a standard exchange format, all attempts will be made to disseminate awareness of the project and its result as widely as possible. A project website will be established as soon as the grant is awarded. The international digital preservation community has well-established and heavily-used communication channels which the project will take advantage of for dissemination. Notification of the start and end of the project along with the website address will be announced in the PADI clearinghouse (<http://www.nla.gov.au/padi/>), the ERPANET preservation website (<http://www.erpanet.org/>), and the Digital Curation Centre Newsroom (<http://www.dcc.ac.uk/news/>). An email press release will be sent to all known preservation-oriented discussion lists.

If accepted by meeting organizers, briefings will be presented at a Coalition for Networked Information Task Force Meeting, at a Digital Library Federation Forum, and at the standards and preservation interest group meetings at the American Library Association annual conference. Articles on the project will be submitted for publication in *D-Lib Magazine* (<http://www.dlib.org>) and the *International Journal of Digital Curation* (<http://www.ijdc.net/>). Finally, participants will submit a paper for acceptance at least one major international digital preservation conference and subsequent publication in the conference proceedings. Possibilities include the IS&T Digital Archiving Conference, the International Digital Curation Conference, and/or the International Conference on Digital Preservation (iPRES).

If the project results in a new METS profile, it will be registered and linked to from the Profiles page of the METS website. Any recommendations affecting PREMIS will be posted to the PREMIS Maintenance Activity pages maintained by the Library of Congress.

6. Sustainability

According the NLG Grant Program Guidelines, criterion for the evaluation of sustainability is "the extent to which the project's benefits will continue beyond the grant period."

Two expected outcomes of this project are a set of recommendations for improvements to the PREMIS Data Dictionary and/or schema, and a recommendation for a standard information package exchange format for preservation repositories. Both of these should have a long-lived impact on the preservation landscape.

The PREMIS recommendations will go to the PREMIS Editorial Committee who, following their usual procedure, will consider these along with changes proposed by

other means for the next upcoming PREMIS revision. The Editorial Committee will accept, reject or modify these proposals after consultation with the PREMIS Implementor's Group, an informal discussion group of international PREMIS users. Accepted changes will become part of the PREMIS standard and should be propagated in implementations worldwide.

The information package exchange format for repositories will be made available to the community for further testing, experimentation and refinement. We expect it to evolve into a *de facto* standard. Potentially it could contribute towards a formal ISO standard associated with the OAIS family of archiving standards.

An immediate impact of the project will be the incorporation of the ability to export and ingest information transfer packages into the open source DAITSS preservation repository application. Changes made to aDORe and DSpace will also be available for other institutions to use. In addition, an anticipated long-term impact of the project is that other repository systems, both open source and commercial, will be enhanced to offer this capacity.

REFERENCES

- [1] Abby Smith, "Distributed Preservation in a National Context: NDIIPP at Mid-point," *D-Lib Magazine* 12 no.6 (June 2006), www.dlib.org/dlib/june06/smith/06smith.html.
- [2] Adam Farquhar and Helen Hockx-Yu, "Planets: Integrated Services for Digital Preservation," *International Journal of Digital Curation* 2 no. 2 (2007), www.ijdc.net/ijdc/article/view/46.
- [3] Consultative Committee on Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, January 2002, public.ccsds.org/publications/archive/650x0b1.pdf.
- [4] Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC), February 2007, www.crl.edu/PDF/trac.pdf.
- [5] It's About Time: Research Challenges in Digital Archiving and Long-term Preservation, (April 2002), p xv, www.digitalpreservation.gov/library/pdf/NSF.pdf.
- [6] *D-Lib Magazine*, 11 no.12 (December 2005), www.dlib.org/dlib/december05/12contents.html.
- [7] Clay Shirky, *Library of Congress Archive Ingest and Handling Test (AIHT) Final Report*, June 2005, www.digitalpreservation.gov/library/pdf/ndiipp_aiht_final_report.pdf.
- [8] MathArc: Ensuring Access to Mathematics Over Time website, www.library.cornell.edu/dlit/MathArc/web/index.html.

[9] IMLS website, www.imls.gov/applicants/basics.shtm.

University of Florida, Florida Center for Library Automation
Towards Interoperable Preservation Repositories (TIPR): A Demonstration Project

SCHEDULE OF COMPLETION

This schedule of completion shows all primary project activities and the amount of IMLS funding devoted to each. Since it does not include cost-sharing contributions, it does not give a true representation of the total amount of resource going into each activity. It does give a true picture of the expected timing and duration of project activities.

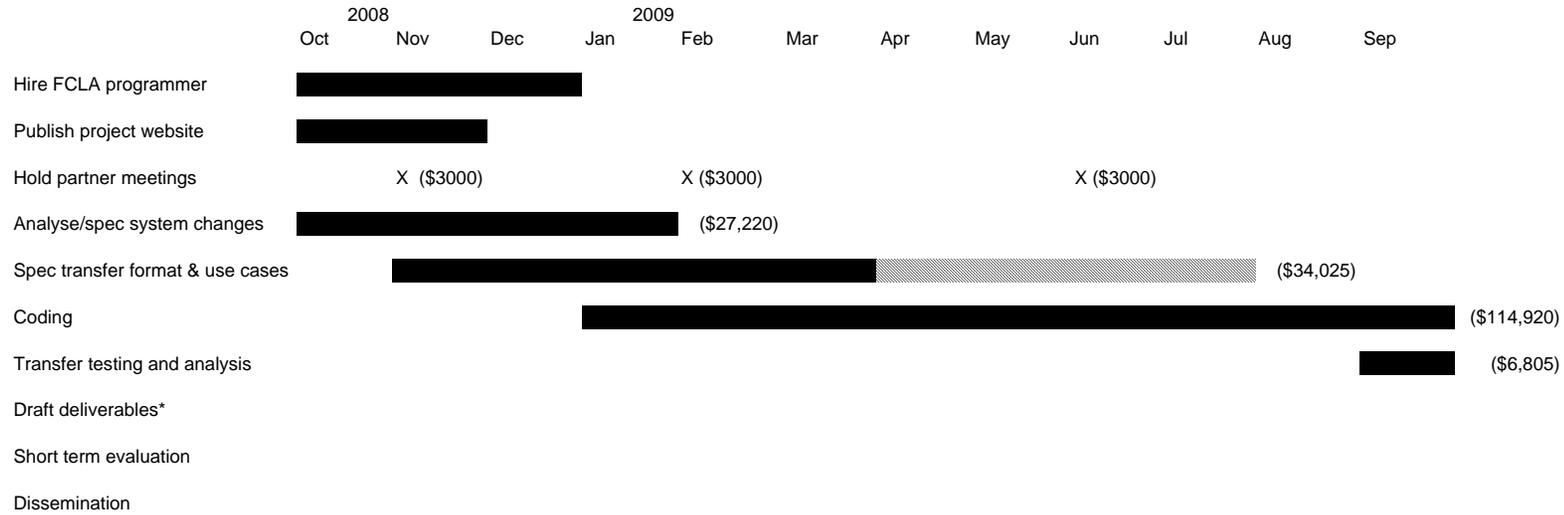
Solid black lines show the time allotted to focusing on each activity. Lighter lines like this:  show periods of time when the activity may be ongoing as a minor activity depending on how coding and testing go.

The amount of IMLS funding devoted to each activity in a year is shown in parenthesis at the end of the line for that activity. Only funding for direct costs is shown.

The total amount of IMLS funding represented for each year does not match the amount shown in the budget because of the omission of two items: 1) the \$4000 allocated to travel to IMLS-designated meetings each year is not shown, because we don't know at this time when those meetings will occur; 2) the IMLS-funded indirect costs awarded to partner organizations is not shown (since only direct costs are shown). To make each year agree with the yearly budget, add in \$4000 travel and partner indirect costs as follows:

Year1:		Year 2:	
Total on schedule	\$191,973	Total on schedule	\$109,095
IMLS travel	\$4,000	IMLS travel	\$4,000
Partner IDC	\$41,345	Partner IDC	\$39,877
Total funding	\$237,316	Total funding	\$152,972

PROJECT YEAR 1



PROJECT YEAR 2

