

Towards Interoperable Preservation Repositories (TIPR)

Priscilla Caplan

Florida Center for Library Automation

5830 NW 39th Avenue

Gainesville, FL 32606

(352) 392-9020 x324

pcaplan@ufl.edu

ABSTRACT

The TIPR (Towards Interoperable Preservation Repositories) project is a partnership between the Florida Center for Library Automation, Cornell University Library, and New York University, funded for two years by the Institute of Museum and Library Services (IMLS). TIPR is based on the assumption that responsibility for long-term digital preservation must be distributed across a number of stewardship organizations running heterogeneous and geographically dispersed 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 a common, standards-based transfer format capable of transporting rich preservation metadata as well as digital objects, and repository systems must be capable of exporting and importing information packages utilizing this format.

The project, which is reaching the midpoint of its second year, has drafted, implemented, and tested a specification for a Repository Exchange Package (RXP), a hierarchical packaging format designed to facilitate the exchange of Archival Information Packages (AIPs) between digital repositories. The RXP encodes structural and preservation metadata using METS and PREMIS, two widely used schema in the cultural heritage community. It is agnostic to the application software used by the sending or receiving repositories or the number of representations included in any AIP.

Categories and Subject Descriptors

J.m [Computer applications]: Miscellaneous

General Terms

Standardization

Keywords

Digital preservation, Preservation repositories, Repository interoperability, RXP, TIPR

1. BACKGROUND

It has long been axiomatic that responsibility for long-term digital preservation must be distributed across a number of stewardship organizations running heterogeneous and geographically dispersed repositories. This is a fundamental principle of all major national and international preservation initiatives, including the U.S. National Digital Information Infrastructure and

Preservation Program (NDIPP), the UK Digital Curation Centre, and Digital Preservation Europe. As Neil Beagrie wrote in a 2002 survey of national digital preservation initiatives, "In none of the countries surveyed is there a single national initiative for digital preservation. Rather, there are many institutional missions that are being extended into the digital domain, including those of national institutions such as the national archives and national libraries." [1]

Distributed digital preservation is a necessity because no one institution, no matter how large or well-funded, has the resources or expertise to preserve more than a small fraction of the world's output of commercial and non-commercial publications, performances and data. It is also desirable, because a network of cooperating preservation repositories offers the community as a whole many advantages. Repositories can specialize in particular genres of materials, in particular preservation strategies, and/or in particular file formats. Important content can be archived in multiple places, improving its chances for long-term accessibility. The community can experiment with different funding models and charging algorithms, different forms of governance and participation, and different service models. Despite the huge strides made in the field over the last ten years, we have to remember that digital preservation as a domain of research and practice is still in its infancy, and can only benefit from the blossoming of a hundred flowers.

At the same time, a landscape composed of geographically distributed, organizationally diverse and technically heterogeneous repositories presents its own challenges, among them the need to transfer archived content from one repository to another safely and efficiently. There are many use cases for repository-to-repository transfer, but the TIPR project has focused on three.

The first is succession planning. The TRAC checklist for certifying trustworthy repositories requires as its second criterion (A1.2) that the repository have "an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope." [2] The standard goes on to advise that a formal succession plan should include the identification of one or more trusted inheritors, if applicable.

For many working repositories including the Florida Digital Archive, the succession plan preferred by the stakeholders would be to transfer responsibility for the archived content to a similar repository. Such an arrangement would require the resolution of a

number of difficult political, financial, legal and technical issues, but its very feasibility depends on the ability to transfer content from one repository to another without loss of information critical to its preservation.

A second use case concerns the best practice that important content should be archived in more than one repository, preferably using different repository applications, different hardware and system software, and different approaches to preservation. Just as wise investors diversify their assets and hedge their investments, those responsible for important scientific, business or cultural heritage materials will want to do the same. It seems likely that as the field of digital preservation matures, more repository options will emerge including some offering highly specialized treatment of certain types of materials. Content owners will want to take advantage of these new opportunities, even if their content is already archived elsewhere. Like the succession case above, this requires the ability to transfer content from one repository to another without loss of critical preservation information.

A third and final use case addresses the fact that no system lasts forever, and repository managers will inevitably have to face the need to upgrade the application they are running or replace it with another. Librarians in particular know what advantages a common, standard transfer format offers in system migration, as their MARC bibliographic records can be exported and imported by any system marketed to libraries.

Towards Interoperable Preservation Repositories (TIPR) is a partnership between the Florida Center for Library Automation, Cornell University Library, and New York University, funded for two years by the Institute of Museum and Library Services (IMLS). The project was initiated in response to the real and perceived, current and future, need of those responsible for digital preservation repositories to exchange archived content. The project asserts that practical repository-to-repository transfer of archived content requires a common, standards-based transfer format capable of transporting rich preservation metadata as well as digital objects, and further asserts that repository systems must be capable of exporting, importing, and to some extent, understanding information packages utilizing this format.

2. TECHNOLOGY

In the United States, "distributed digital preservation" has become almost synonymous with private LOCKSS networks (PLN). So much has been written about PLNs that we will not discuss them here, except to say that they are not relevant to this problem space because they don't address the use cases described above. Although PLNs can provide for the geographical distribution of content, by definition all sites within a network run identical software and very similar hardware, so they provide no architectural heterogeneity. A PLN could conceivably be designated as a successor repository, but the LOCKSS application does not create, store or use detailed preservation or format-specific technical metadata, so it is unlikely it would be chosen to inherit content from a more fully featured repository system. And PLNs only replicate content, they do not import it from or export

it to external repositories, so no standard transfer format is required. That said, a PLN taken as a whole could serve as an exchange partner within the TIPR model, a scenario that could be tested in the future.

Technologies that do address this problem space include the Echo Depository's Hub and Spokes architecture (HandS) and the Open Archives Initiative's Object Reuse and Exchange specification (OAI-ORE).

The Echo Depository was an NDIIPP-funded project that ran from 2004-2009. It included a research thread to evaluate and compare commonly used institutional and preservation repository systems. This analysis ascertained that the systems examined had minimal interoperability and little support for active preservation strategies such as format migration. The HandS architecture is designed to compensate for both of these deficiencies. In this model each individual repository system is a relatively dumb spoke communicating with a smart central Hub. The central Hub service can pull a package from a source repository, enrich it with metadata, map it to a common exchange format, remap from the exchange format to the native format of a target repository, and push the reformatted, enriched package into the target repository. The exchange format defined by HandS is based on MODS, METS and PREMIS standards, and uses applicable standards for format-specific technical metadata. Metadata not included in the source package is added by the Hub.[3]

The TIPR project has from the beginning focused only on the practical requirements and constraints affecting exchange among digital preservation repositories in the real world. The TIPR model differs from the HandS model in that it does not require a smart central Hub. In a grant-funded project, a Hub is feasible because the project has the motivation, resources and expertise to build and maintain it. In a post-project, operational environment, finding the means to support a Hub would be a financial and organizational challenge. In TIPR, all exchanges are peer-to-peer, requiring the developers of each repository software application to support an export to and import from the common exchange format. (See Figure 1.) The investment required is reasonably small, and it falls upon parties with both the expertise and motivation to provide this functionality.

OAI-ORE is a set of standards for the description and exchange of aggregations of Web resources. Aggregations are described by resource maps which can be serialized according to any number of XML syntaxes including RDF and Atom. Aggregations can in turn be included as resources in larger aggregations. Since archived content can be seen as a hierarchy of aggregations, OAI-ORE is in some ways a natural model for data exchange among disparate repositories. ORE was used to successfully exchange content between two institutional repository systems, Fedora and DSpace, in a proof-of-concept exercise.[4] Although these are not preservation repositories, it is possible to extrapolate to a preservation context.

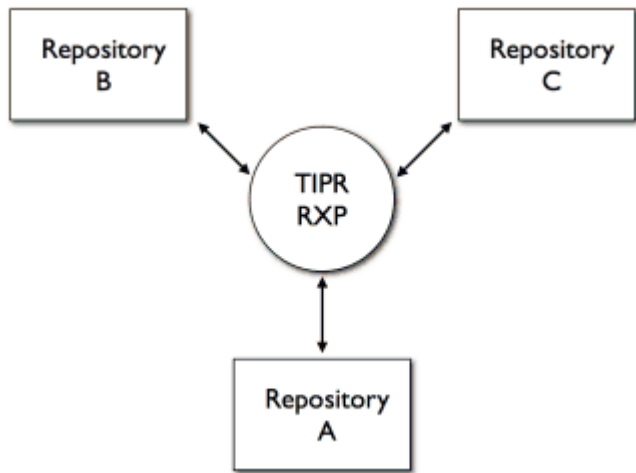


Figure 1: The TIPR model of exchange.

The TIPR model differs from OAI-ORE in that it is oriented specifically towards exchange among digital preservation repositories. There is no requirement that the content transferred be Web resources or that the underlying transport protocol be HTTP; TIPR is equally usable by a dark archive sending content on tape via a courier service. Moreover, TIPR is cognizant of the need for an unbroken chain of digital provenance to support the continuing authenticity of the content, and expects operating preservation repositories to understand both the syntax and semantics of preservation metadata documenting this.

The TIPR model is based on concepts from the Open Archival Information System Reference Model and the PREMIS Data Dictionary for preservation metadata. [5, 6] In the OAIS framework, the open archival information system (preservation repository) ingests, stores and disseminates information packages that consist of content data files to be preserved and metadata describing them. A Submission Information Package (SIP) submitted by a producer to an archive is transformed into an Archival Information Package (AIP) for archival storage. The AIP in turn can be transformed into a Dissemination Information Package for delivery to a consumer.

PREMIS is a standard for preservation metadata that includes an object model for what types of things should be described. These include bitstreams, files, and aggregations of files called "representations," which are defined as the complete set of files needed to render a particular intellectual entity. So, for example, if the intellectual entity is instantiated by a web page made up of an HTML file and a GIF image file, the two files together constitute a representation of that intellectual entity. If a repository creates a new version of the image in PNG format, it has created a second representation of the same intellectual entity, one consisting of an HTML file and a PNG file. Depending on the preservation strategy employed by the repository, it may

retain one or both representations. If it retains both, they may be considered part of the same AIP or may be different but related AIPs.

TIPR defines a common export format called the Repository Exchange Package (RXP) based on the Metadata Exchange and Transport Standard (METS). The point of the RXP is to allow heterogeneous repositories to exchange packages with minimum loss and maximum understanding. TIPR assumes that the basic unit of transfer is an AIP disseminated from the source repository as a DIP, regardless of how the repository defines an AIP. An RXP can contain one or more representations, but exactly one must be identified as the "active" representation. This information can be used by the receiving repository, which may implement a different preservation protocol, to deconstruct the incoming RXP appropriately.

RXP Minimal Structure

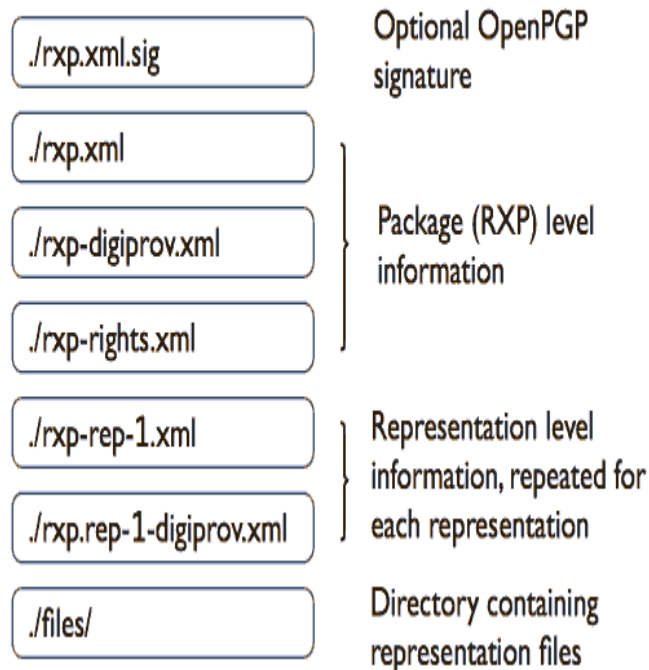


Figure 2: A minimal RXP.

Figure 2 illustrates the structure of an RXP. Three package level XML documents describe the exchange package as a whole. *rxp.xml* is a METS document identifying the package, the sending repository, and the active representation. *rxp-digiprov.xml* is a PREMIS document containing the digital provenance of the package. *rxp-rights.xml* is a PREMIS document with package level Rights information.

Following the package-level information, each representation in the AIP must be described by a pair of documents, a METS document minimally containing a filesec and structMap, and a PREMIS document minimally containing digital provenance (event) information for the representation. These are pointed to in the *rxp.xml* document by mdRef elements.

Finally, content files (that is, the actual contents of the source AIP/DIP) are located in a directory called /files. Optionally, a digital signature can be generated over *rxp.xml* and stored as *rxp.xml.sig*.

The specification is being exercised and refined by a series of transfer tests carried out by the partner institutions. Each of the three partners operates a working preservation repository that has archived digital content in production. Each repository implements or can implement a transformation-based preservation protocol, and each supports detailed preservation-related metadata. At this time testing is ongoing, but so far it has indicated the following:

a) The effort required to create an RXP is quite reasonable. It requires no code change to a repository application, as it can be done by transforming the native DIP after dissemination.

b) It has been more difficult, but not prohibitively difficult, for the partner repositories to ingest a foreign RXP. Code changes were required because the partner's repository systems were designed to create and store digital provenance information for packages at ingest, but they did not expect pre-existing provenance to be included in a new SIP.

c) Maintaining a continuous record of ownership is tricky when a package is transferred more than once; that is, from site A to site B to site C. Package level provenance is contained in *rxp-digiprov.xml*, so the receiving repository can maintain provenance by storing that document as an archived content file. File level provenance is more complex. The project will publish guidelines and examples for handling this.

d) The fact that a repository can ingest a package transferred as an RXP from a source repository successfully does not mean that the use cases of diversification, succession, and system migration are served. Each use case carries its own demands for what must be maintained and/or understood by the receiving repository.

An important lesson is that a common transfer mechanism is only a part of the infrastructure needed for these real world use cases to be satisfied. Repository managers will have to establish inter-repository agreements similar to service level agreements spelling out requirements related to ownership and access, preservation treatment of transferred content, the level of acknowledgement and reporting the source repository can expect, and many other issues of policy and practice.

The TIPR project will continue through the end of September, 2010. The draft RXP specification is available on the project website (wiki.fcla.edu:8000/TIPR) as are shell scripts and Schematron schemas to validate RXP XML documents according to the specification. Additional documentation and examples will be made available by the end of the project.

3. IMPACT AND FUTURE PLANS

The use cases indicate that the digital preservation community has a real and currently unmet need for a uniform and practical mechanism for repository-to-repository transfer. The potential impact of a widely used community standard for repository-to-repository exchange of archival information packages is clearly quite significant.

TIPR project partners hope to see some take-up and early adoption of the RXP by developers and implementers of commonly-used open source and commercial preservation applications. Further testing in a broader context should lead to further improvement of the specification, and to consensus validation of the underlying principles that the repository-enriched metadata pertaining to a stored AIP must accompany that AIP in transfer, and that an unbroken record of digital provenance must be maintained through any transfer.

4. REFERENCES

- [1] Bernstein, A. 2009. Commercial companies have their say. *Planetarium* 8 (Dec. 2009), 6.
- [2] OCLC and CRL. 2007. Trustworthy repositories audit & certification: criteria and checklist (TRAC).
- [3] Habing, T., et. al. 2009. Developments in Digital Preservation at the University of Illinois: The Hub and Spoke Architecture for Supporting Repository Interoperability and Emerging Preservation Standards. *Library Trends* 57,3 (Winter 2009), pp. 556-579. DOI = DOI: 10.1353/lib.0.0052
- [4] Tarrant, D., et. al. 2009. Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications. *Code4Lib Journal* 6 (Mar. 2009).
- [5] Consultative Committee on Space Data Systems. (2002). Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1: Blue Book. Issue 1.
- [6] PREMIS Data Dictionary for Preservation Metadata, version 2.0, March 2008.