Title:
Towards Interoperable Preservation Repositories (TIPR)

Author:
Caplan, Priscilla, Florida Center for Library Automation
Kehoe, William, Cornell University
Pawletko, Joseph, New York University

Publication Date:
10-05-2009

Series:
iPRES 2009: the Sixth International Conference on Preservation of Digital Objects

Publication Info:
iPRES 2009: the Sixth International Conference on Preservation of Digital Objects, California Digital Library, UC Office of the President

Permalink:
http://www.escholarship.org/uc/item/5wf5g5kh

Multimedia URL:
http://www.cdlib.org/services/uc3/iPres/video.html?file=ipres/Pawletko&title=Priscilla%20Caplan%20and%20Joseph%20Pawletko%3A%20TIPR's%20Progress%20%28Towards%20Interoperable%20Preservation%20Repositories%29

Abstract:
TIPR, Towards Interoperable Preservation Repositories, is a project funded by the Institute of Museum and Library Services to create and test the Repository eXchange Package (RXP). The package will make it possible to transfer complex digital objects between dissimilar preservation repositories. For reasons of redundancy, succession planning and software migration, such repositories must be able to exchange copies of archival information packages with each other. Every different repository design, however, describes and structures its archival packages differently. Therefore each type produces dissemination packages that are rarely understandable or usable as submission packages by other repositories. The RXP is an answer to that mismatch. Other solutions for transferring packages between repositories focus either on transfers between repositories of the same type, such as DSpace-to-DSpace transfers, or on processes that translate a specific dissemination format into a specific submission package. Rather than build translators between many dissimilar repository types, the TIPR project has defined a standards-based package of metadata files that can act as an intermediary information package, the RXP, a lingua franca all repositories can read and write. In this paper we present the assumptions and principles underlying the TIPR concept of repository-to-repository exchange, and proceed to describe three

aspects of the TIPR project: the RXP format itself; the tests we are conducting to prove and improve the use of the RXP; and finally, issues that have arisen in the course of the project so far.

iPRES2009

THE SIXTH INTERNATIONAL CONFERENCE ON THE PRESERVATION OF DIGITAL OBJECTS

Proceedings

October 5-6, 2009
Mission Bay Conference Center
San Francisco, California

California Digital Library

Towards Interoperable Preservation Repositories (TIPR)

Priscilla Caplan
Ass't Director for Digital Library Services
Florida Center for Library Automation
5830 NW 39th Avenue,
Gainesville, FL 32606
pcaplan@ufl.edu

William Kehoe
Enduring Access Analyst
Cornell University Library
Ithaca NY 14853
wrk1@cornell.edu

Joseph Pawletko
Software Systems Architect/Technical Lead
Bobst Library, New York University
70 Washington Square South
New York, NY 10012
jgp@nyu.edu

Abstract

TIPR, Towards Interoperable Preservation Repositories, is a project funded by the Institute of Museum and Library Services to create and test the Repository eXchange Package (RXP). The package will make it possible to transfer complex digital objects between dissimilar preservation repositories. For reasons of redundancy, succession planning and software migration, such repositories must be able to exchange copies of archival information packages with each other. Every different repository design, however, describes and structures its archival packages differently. Therefore each type produces dissemination packages that are rarely understandable or usable as submission packages by other repositories. The RXP is an answer to that mismatch. Other solutions for transferring packages between repositories focus either on transfers between repositories of the same type, such as DSpace-to-DSpace transfers, or on processes that translate a specific dissemination format into a specific submission package. Rather than build translators between many dissimilar repository types, the TIPR project has defined a standards-based package of metadata files that can act as an intermediary information package, the RXP, a lingua franca all repositories can read and write.

In this paper we present the assumptions and principles underlying the TIPR concept of repository-to-repository exchange, and proceed to describe three aspects of the TIPR project: the RXP format itself; the tests we are conducting to prove and improve the use of the RXP; and finally, issues that have arisen in the course of the project so far.

Introduction

Towards Interoperable Preservation Repositories (TIPR) is a two-year project partnership between the Florida Center for Library Automation, Cornell University and New York University, funded by the Institute of Museum and Library Services (IMLS). The goal of the project is to develop, test and promote a standard interchange format for exchanging stored information packages among OAIS-based preservation repositories.

The use cases for transferring copies of stored information packages from one repository to another are entirely practical. For example, at this time there are few true preservation repositories and most are operated for the use of particular constituencies. Imagine a library that archives its digital content in the only repository available to it, say a university-operated institutional repository. A few years later a new preservation repository specifically tailored to Geographic Information Systems opens for business. The library may want to collect its archived GIS content from the institutional repository and deposit a copy in the special GIS repository.

In a second use case, the institutional repository posited in the first case ceases operation, perhaps because the university has an opportunity to become a member of a larger shared repository with more preservation functionality. In this case the entire stored content of the institutional repository must be transferred to the shared repository.

In a third case, the university might decide to change repository systems from the old institutional repository it was running to a second-generation preservation repository system. In this case again the entire content of the institutional repository must be transferred to the new preservation application.

In the OAIS model, digital objects are submitted to preservation repositories as Submission Information Packages (SIPs); the process of Ingest transforms a SIP into an Archival Information Package (AIP) for storage; and the process of dissemination transforms an AIP into a Dissemination Information Package (DIP) for export. In OAIS terms, then, transferring a copy of a stored object from repository A to repository B is a matter of A transforming an AIP into a DIP to be ingested as a SIP by B. Because different repository systems describe and structure their archival packages differently, a DIP produced by one repository is unlikely to be directly usable as a SIP by another. Therefore A's DIP must be somehow be transformed into a SIP that B can ingest.

The TIPR approach to this problem is to define a common exchange package format, the Repository Exchange Package (RXP). In this model, every repository need only understand two package formats, the RXP and its own native DIP, in order to be able to exchange packages with all other repositories. The TIPR model is shown in Figure 1. (An alternate model called Hub and Spokes or "HandS" is used by the Echo Depository project, in which a central Hub application performs the translations on both sides.)

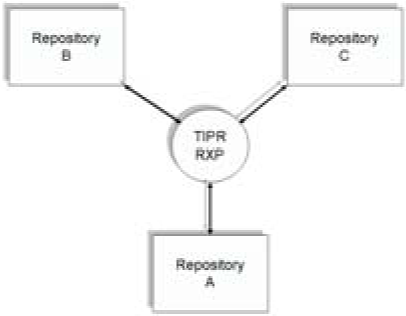

Figure 1: The TIPR model of package exchange

Design Issues

The RXP was designed to meet the following requirements:

1) The exchange package must use well known and accepted standards in the cultural heritage preservation community. The project did not want to burden the community with yet another "standard" conflicting with or overlapping with standards already in use.
2) The exchange package must be flexible enough to accommodate any repository's AIP; that is, it must be agnostic to the internal structure of the AIP. 3) The exchange package must contain enough information for the target repository to know what it is receiving both at the package level and the representation level.
3) Some selected information provided by the sending repository must actually be understood by the receiving repository.

Because of requirement (1), the TIPR RXP is based upon METS and PREMIS. TIPR assumes that the preservation community knows (or ought to know) how to interpret METS syntax and PREMIS semantics. These two standards represent the core of a meaningful exchange.

Requirements (2) and (3) mean that any repository system can use the RXP no matter what their treatment of representations with no change to their internal AIP architectures. In the PREMIS data model, a representation is defined as the set of files needed to fully render an intellectual entity. For example, a particular journal article might consist of 14 data files: an XML file, a stylesheet, and a dozen images. An alternate representation of the same article might consist of a single PDF file. Depending on the repository architecture, an archive could treat the two representations as a single AIP or two AIPs. If the images were migrated from JPEG to JPEG2000, a third representation would be created, and again, a given repository could treat the three representations as one, two or three AIPs. The three TIPR project partners have implemented preservation repositories with quite different approaches to representations, so they provide a good testbed for these requirements.

Requirement (4) is possibly the heart of the project and what distinguishes TIPR from other exercises in package exchange. The TIPR partners believe that package transfer is not a matter of the mere duplication or replication of data. Certain information critical to digital preservation must be not only stored but also understood by the target (receiving) repository. Understanding in this context means that the metadata elements and values can be mapped to equivalent elements and values in the receiving system. A meaningful exchange in the preservation context dictates that interoperability must be semantic as well as syntactic.

A major early task of the project was to decide which types of metadata potentially maintained by the sending repository would need to be understood by the receiving repository. Various categories of administrative, preservation, and format-specific technical metadata were analyzed in turn. The project decided that most types of metadata could be recreated by the receiving repository, simply stored as received, or covered by a repository-to-repository Service Agreement. However, information pertaining to rights and to digital provenance (the history of ownership and actions affecting the object) must be understood. The case for rights metadata is straightforward, since actionable rights information may control what access to the object is allowed and what preservation actions can be performed.

The case for digital provenance information deserves some elaboration. An OAIS-based preservation repository will perform many actions on a SIP in order to transform it into an AIP, which may or may not include creating transformed versions of source files. A common preservation strategy for both libraries and archives is to guard against format obsolescence by reformatting archived content files. A normalized version of a source file may be created in a format considered to be more preservation-worthy (stable, well understood, nonproprietary, etc.). A migrated version may be created in a more current version of the format, or a successor format.

The original source file may be retained in archival storage or discarded in favor of the derivative version(s). In a preservation environment in which transformations may occur, the only way to guarantee the continued authenticity of the digital object is to maintain an unbroken record of digital provenance.

In PREMIS, rights and permissions relevant to the preservation of the object are described by the Rights entity, and digital provenance is described by Events. Both Rights and Events can be associated with Agents, which can be persons, organizations or software. TIPR uses PREMIS as a meta-language for expressing these concepts regardless of the way they are represented in the sending or receiving repository system.

The Repository Exchange Package (RXP)

A minimal RXP consists of exactly five required XML metadata files and a directory of files from the sending repository's AIP. Additional files may be included as shown in Figure 2.

The three files rxp.xml, rxp-digiprov.xml and rxprights.xml contain information about the exchange package itself. rxp.xml is a METS document identifying the package and the sending repository. The METS root element must have an OBJID attribute with an info:uri that is unique to the sender, and the METS header must have an AGENT attribute identifying the sender. It uses METS mdRef elements to point to three (or more) of the remaining XML files defined in the RXP specification. The only metadata files not referenced directly by mdRef elements are those which contain representation-level digital provenance (rxp-rep-n-digiprov.xml, described below).

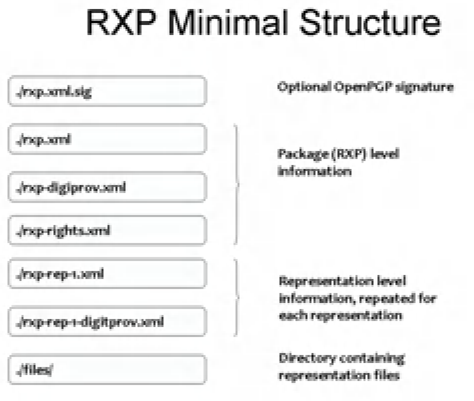
RXP Minimal Structure
./rxp.xml.sig
Optional OpenPGP signature
./rxp.xml
./rxp-digiprov.xml
./rxp-rights.xml
Package (RXP) level information
./rxp-rep-1.xml
./rxp-rep-1-digitprov.xml
Representation level information, repeated for each representation
./files/
Directory containing representation files

Figure 2: RFP minimal structure

rxp-digiprov.xml is a PREMIS document containing digital provenance (Event) information pertaining to the RXP package itself. rxp-rights.xml is a PREMIS document with package level Rights information. Optionally, a fourth package level file may be present, rxp.xml.sig, containing a digital signature in OpenPGP format generated using the sender's private key and rxp.xml.

The remaining two files in Figure 2, rxp-rep-1.xml and rxp-rep-1-digiprov.xml, describe the first representation in the sending repository's AIP. rxp-rep-1.xml is a METS document describing the structure of the representation. rxp-rep-1-digiprov.xml is a PREMIS document containing digital provenance information for the representation. This pair of files should be repeated for every representation n in the AIP, as rxp-rep-n.xml and rxp-rep-n-digiprov.xml.

The files directory contains the files of the representation(s). Any files which must be preserved to fully describe and create the AIP should go in this directory. Whereas the information contained in the RXP metadata files must be understood and preserved, the metadata files themselves need not be retained. All files in the files directory, however, must be preserved as indicated by the Service Agreement between the sending and receiving institutions.

The RXP specification and schema are available on the TIPR project site.

The Transfer Tests

The transfer tests are designed to ascertain the extent to which the partner systems can send and receive packages with minimal loss of information and maximum understanding. The project is unconcerned with the mechanics of file transfer, which would in reality be negotiated between repositories and specified in a Service Agreement. For the purpose of the project, test RXP packages were bundled according to the BagIt specification and transferred by HTTP.

The first step in testing was to make each system capable of outputting a conforming RXP. In most cases partners did not change their repository systems but rather wrote code to transform their native DIP to RXP format. Validation scripts using Schematron (www.schematron.com) were written to validate the resulting RXP XML files.

Each partner then created two RXP format exchange packages, including digital signatures, and sent these to the other partners for ingest. This type of broadcast transfer tests that different AIPs can be transformed into RXPs, and that each receiving repository is capable of transforming an RXP format package into an ingestible SIP.

A third milestone will test a "ring transfer," where repository A sends an RXP to repository B; repository B ingests the packages and exports it as an RXP to repository C; repository C ingests the package and exports it as an RXP to repository A. Repository A ingests the package and compares the resulting AIP to the original AIP in the chain.

Following completion of the ring transfer, the focus in testing will shift to the selection of sufficiently varied source AIPs to exercise the range of issues likely to be encountered in real-world transfer.

As the transfer tests have proceeded, some deficiencies with the RXP exchange format were exposed and rectified. For example, initially the partners preserved RXP identifiers by using the OBJID and LABEL attributes in the top METS element of rxp.xml. This was found to be insufficient for preserving the history of exchange if the same package was transferred to two or more institutions (e.g., from repository A to B to C). This led to the definition of rxp-digiprov.xml which can now track the full history of a particular RXP.

Each institution's repository had its own method for identifying objects, events, and agents. To avoid potential conflicts between these identifiers when transferring digital provenance, the RXP specification was amended to require all PREMIS identifiers to be info:uris. This requirement makes identifiers universally exchangeable between any number of repositories.

Other technical issues that arose from the first transfer test included:

- complications exchanging public keys (FCLA gave multiple keys, one for use in rxp.xml.sig and one for exchange, which confused NYU and CUL)
- some complications posting data to a public HTTP site (NYU had a large package broken into many parts)
- bags were hassle free, but CUL's tools reported that FCLA bags were missing an optional fixity value
- size limitations for large data files required some patches to certain FCLA services

Transfer Issues

Some longer-term issues have also arisen in testing so far. One exposes a limitation of the project's reliance on PREMIS. Receiving repositories need some description of the exchange package itself, including its own digital provenance and what high level rights adhere to it. PREMIS is capable of describing this, but the highest level of description in PREMIS is a representation object. An RXP package can contain multiple representations, and is more comparable to an Intellectual Entity than to anything else in the PREMIS data model. Unfortunately, the Intellectual Entity is out of scope for PREMIS as currently conceived. The TIPR project has requested that the PREMIS Editorial Committee consider allowing PREMIS elements describe intellectual entities when applicable. In the meantime, the project is using PREMIS anyway, in violation of that standard.

A second issue concerns limits on what an exchange package can reasonably be expected to communicate. It was evident very early that successful exchange will require an agreement between participating repositories to supplement the information contained in the RXP. At a minimum, the service agreement must document the following:

1) details of RXP composition by the source repository in this particular transfer, where the RXP specification allows options;
2) how the RXP will be transferred from source to target repository;
3) actions to be performed by the target repository on receipt of the RXP;
4) rights and permissions agreed upon by the source and target repositories;
5) archiving and preservation treatment of the ingested RXP by the target repository;
6) financial arrangements between source and target repositories;
7) legal aspects of the arrangement.

The first three items are quasi-technical, assuring the mechanics of transfer are addressed. Stipulations concerning (3), for example, would include what acknowledgement the sending repository can expect to get (if any) at the time the RXP is received by the target repository and when it is ingested. Item (4) is necessary because the PREMIS Rights entity lacks the expressiveness required to describe some complex rights without referencing external documents. At this time, a service agreement should detail how these rights are to be interpreted. In the future, as PREMIS rights become more expressive, a deeper exploration of rights will be necessary. The last three items are of critical concern to the owners of the content, and might influence decisions like which repository(ies) to designate in succession planning.

Acknowledgements

The TIPR project is supported by a two-year grant from the Institute of Museum and Library Services. The authors thank Francesco Lazzarino and Marly Wilson for their help in drafting this paper.

References

Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System. January 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf

Hub and Spoke Framework Tool Suite. http://dli.grainger.uiuc.edu/echodep/hands/index.html

Metadata Encoding and Transmission Standard (METS) official website. http://www.loc.gov/standards/mets/

PREMIS Editorial Committee. March 2008. PREMIS Data Dictionary for Preservation Metadata. Version 2.0. http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

PREMIS Preservation Metadata Schema, version 2.0. 2008 http://www.loc.gov/standards/premis/premis.xsd

TIPR: Towards Interoperable Preservation Repositories Project Website. http://wiki.fcla.edu:8000/TIPR