TOWARDS INTEROPERABLE PRESERVATION REPOSITORIES (TIPR): THE INTER-REPOSITORY SERVICE AGREEMENT

Priscilla Caplan
Florida Center for Library Automation
Gainesville, Florida, USA

William Kehoe
Cornell University Library
Ithaca, New York, USA

Joseph Pawletko
Bobst Library
New York University
New York, New York, USA

ABSTRACT

The TIPR Project (Towards Interoperable Preservation Repositories) runs from October 2008 through September 2010. The aim of the project is to develop, test, and promote a standard format for exchanging information packages among OAIS-based repositories. This paper reviews the use cases for the transfer of information from one repository to another, reviews the Repository eXchange Format (RXP) developed by TIPR, and discusses the need for additional information not contained in the exchange package itself. It looks at two existing specifications, the Producer-Archive Interface Methodology and Into the Archive (Wege ins Archiv), in the context of inter-repository transfer. Finally it outlines information required in an inter-repository service agreement.

1. INTRODUCTION

The TIPR Project (Towards Interoperable Preservation Repositories) was begun in October 2008 with the aim of developing, testing, and promoting a standard format for exchanging information packages among OAIS-based preservation repositories. The project was premised on the idea that there are at least three real-world use cases requiring one repository to transfer an archived AIP for ingest into a different repository system:

- diversification (the owners of valuable content want it stored in multiple, heterogeneous repositories)
- succession (the source repository is ceasing operations and transferring its content to one or more other repositories)
- system migration (the repository is replacing its applications software and must migrate its archived content to the new system)

Over the past two years, the project participants have drafted and tested a package format, the Repository Exchange Package (RXP), designed to facilitate the transfer of an AIP from one repository to another. Based on the METS and PREMIS standards, the RXP describes the provenance and structure of one or more versions of a digital object.

Our prior experiences using METS and PREMIS influenced us to adopt a design philosophy for the RXP that favors constraint over flexibility. We had found that local and optional metadata elements often hinder interoperability by making exchange more difficult, impeding semantic understanding, and/or rendering the data less useful in the target systems. However, in the real world repositories are based on different software applications and run by different institutions, and there is little consistency in data models or metadata.

The TIPR approach to this dilemma is to constrain the METS and PREMIS elements in the RXP and, at the same time, to complement that constraint with some allowable flexibility, embodied in an inter-repository service agreement. The agreement complements the RXP by expressing each organization's intentions and responsibilities. The RXP bears the constrained metadata for machine transfer, while the inter-repository service agreement makes local conditions explicit, and can vary according to the circumstances and use case for any given transfer. As such, the inter-repository service agreement can be seen as a form of submission agreement between a producer and an archive.

In the next section we review the structure and content of the RXP. Section 3 reviews two specifications for the transfer of information to a digital preservation repository. In section 4 we explore the applicability of these specifications to the case of inter-repository transfer. Section 5 looks at the information required in an inter-repository service agreement.

2. A BRIEF LOOK AT THE REPOSITORY EXCHANGE PACKAGE

Conceptually, the RXP consists of three sets of files: 1) the component files of the digital object(s) being transferred; 2) metadata files describing the structure and provenance of these files; and 3) metadata files describing the structure and provenance of the package itself. Structure is described in METS documents, provenance is encoded in files containing PREMIS elements, and the digital object component files are bundled in a flat directory, their original relationships described in the METS document.

More than one version of a digital object can be packaged in an RXP, each version with its own set of structural and provenance descriptor files. These versions correspond to "representations" in PREMIS terminology.

The RXP is shown schematically in Figure 1, and is described in more detail in [1, 2, 3].

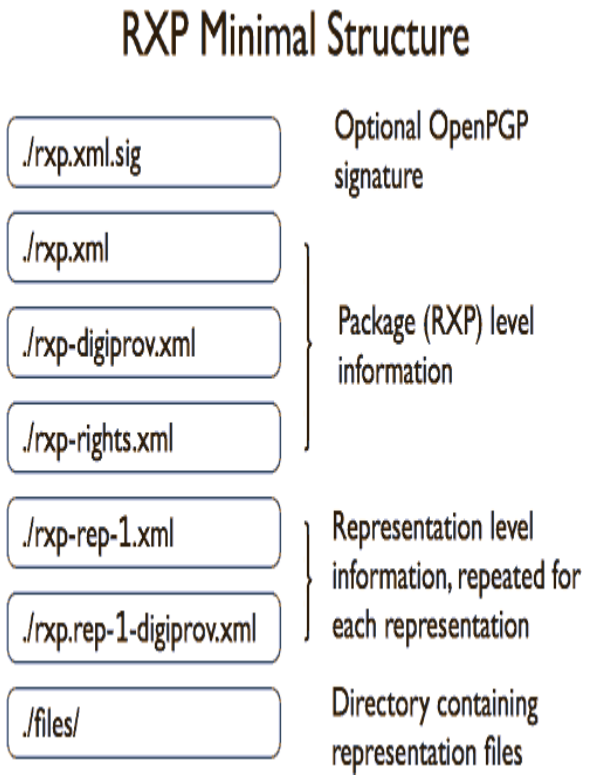

Figure 1: The structure of a Repository Exchange Package (RXP).

3. PRODUCER-ARCHIVE AGREEMENTS

It is well-accepted that the submission of content to a repository for archiving should be governed by a submission agreement. Submission agreements are addressed in the Producer-Archive Interface Methodology Abstract Standard (PAIMAS) [4] and in Into the Archive: A Guide for the information transfer to a digital repository [5].

3.1. PAIMAS

The Producer-Archive Interface Methodology is an ISO standard that builds upon the Reference Model for an Open Archival Information System [reference] and uses terms as defined in that document. Specifically, it elaborates all of the actions and negotiations that a content producer (Producer) and a repository (Archive) must take from their initial contact, through the transmission of SIPS to a repository, to the receipt and validation of the SIPs by the repository. PAIMAS is structured around phases, which must take place in order. A 46-step preliminary phase and a 36-step formal definition phase culminate in the drafting of a mutually acceptable Submission Agreement, after which shorter transfer and validation phases complete the Producer-Archive project.

While PAIMAS specifies in detail a methodology for achieving a Submission Agreement, the actual content of the agreement is largely left to be inferred from the steps leading to its creation. The Submission Agreement is described at a high level as defining the information to be transferred, the transfer process, how SIPs will be validated by the Archive, a schedule for submission, and conditions for changing or breaking the Agreement. Reporting requirements are not listed explicitly, but are implicit in transfer and validation specifications.

3.2. Into the Archive

Into the Archive (Wege ins Archiv, hereafter referred to as the "nestor Guide") is a guide produced by Germany's nestor working group on long-term preservation standards. Its aim is similar to that of PAIMAS, but it is shorter and simpler, and of a more practical than theoretical orientation. Like PAIMAS, the nestor Guide stipulates that the producer and the archive draw up a binding "ingest agreement." Ingest is defined as ending at the point where the archive has received, validated and accepted responsibility for the package, so the scope of the ingest agreement is formally the same as that of the PAIMAS submission agreement.

The nestor Guide is organized around objects, processes, and management, listing practical objectives in these areas and procedures for achieving them. Within this framework, the ingest agreement is simply another objective, rather than the end result of a long process. The ingest agreement covers much the same topics as the Submission Agreement, except that it does not include conditions for modification or termination. It goes beyond the Submission Agreement, however, in including some stipulations about how data is to be treated by the receiving archive. It requires a definition of the significant properties of the objects to be archived, the "technical environment" required for archiving them, and agreed-upon preservation treatment ("migration agreements"). Reporting requirements are not listed as included in the ingest agreement, but reporting is a separate requirement of the information transfer process.

4. REPOSITORY TO REPOSITORY TRANSFER

The case of transfer of an AIP from one repository system to another can be seen as a special case of transfer from producer to archive. It does, however, introduce another set of contextual circumstances and some unique requirements.

4.1. Role of producer.

In both PAIMAS and the nestor Guide, the Producer is formally defined according to OAIS as the party transferring objects to the preservation repository. Both specifications are clear that the Producer does not have to be the original content creator or owner. PAIMAS explicitly allows for a third party to assume the role of

Producer when there is no relationship between the Archive and the true Producer(s), giving the example of a library department entrusted with archiving a collection of CD-ROMS from a number of non-cooperating publishers. Accordingly, in the case of one repository transferring AIPs to a second repository, the sending repository could be considered a proxy producer.

Both specifications, however, carry the implicit assumption that the Producer-Archive relationship is bilateral. In a TIPR-type transfer, the relationship is more likely to be trilateral, although the alignment of players depends on the use case. In the case of diversification, the original producer (the depositor of the AIP held by the sending repository) and the proxy producer (the sending repository) are likely to be equal partners, both communicating with the archive (receiving repository). The case of succession planning may parallel case of diversification, with the original producers playing an active role, or the terminating repository may conduct all negotiations on their behalf. This case is particularly interesting as ingest concludes and in the post-ingest phase, as the receiving repository may now need to maintain relationships with a multiplicity of original producers instead of the single proxy producer (especially when the terminating repository actually ceases to exist).

The case of system migration has parallels to the succession scenario. The sending repository ceases to exist as a repository application, and the receiving repository application takes over the relationship and communications with the original producers. The institutional management of the two repository applications does not change, and of course is the same for each.

4.2. Selection of archive.

PAIMAS posits a protracted period of information exchange between Producer and Archive, at the end of which each side assesses whether or not it is desirable to continue with the project and draft a Submission Agreement. The nestor Guide assumes the two parties have already been determined and information is exchanged only to ensure the appropriate treatment of materials. In a TIPR-type transfer, the different use cases have quite different implications for the selection of a receiving repository. In the case of system migration, the organizational management of the Producer (old repository) and Archive (new repository) can be assumed to be the same, obviating the need for many PAIMAS activities. In the case of succession, the Producer (terminating repository) may not be in a position to undertake many of the steps. Only in the case of diversification are most of the PAIMAS activities likely to apply.

4.3. Selection of content.

In PAIMAS, the selection of content to be preserved is a joint responsibility of the Producer and the Archive to be worked out in the preliminary phase, although the Producer initiates the process by describing the type of information it wants to preserve. In the nestor Guide, the final selection of content falls to the archive. The assumed context is traditionally archival, where a government agency or institution exposes its entire collection to the repository, which has a legal or contractual mandate to assume responsibility for items which meet certain criteria.

In a TIPR-type transfer, the three use cases have different implications for selection. In the case of diversification, it is almost certainly the original or proxy Producer who will identify specific content and seek a repository most capable of preserving it. In the case of system migration, there is likely to be no selection at all, the assumption being that all of the content in the old system will be transferred to the new. In the case of succession, either all of the terminating repository's content will be transferred to a single receiving repository, or variously defined subsets of content (for example, by media type, or by original owner) will be identified for transfer to different repositories. While the receiving repository will have some say in what it will agree to take, in no case does it have primary responsibility for selection. In this respect PAIMAS models selection better than the nestor Guide.

4.4. SIP creation.

Both PAIMAS and the nestor Guide assume the Producer is creating an original SIP (i.e., a SIP for first-time archiving). In the case of repository to repository transfer of a SIP created from a previously archived AIP, the sending repository has additional constraints; for example, it may not be able to obtain additional metadata from the original producer. At the same time, the sending repository is likely to have enriched the original AIP with metadata of its own, such as format-specific details, validation results, and processing history. While these factors will complicate the negotiation of a transfer project, the existence of a standard transfer format such as the RXP dramatically simplifies and/or obviates the need for a number of steps defined in PAIMAS.

4.5. Role of agreement.

In the nestor Guide, the ingest agreement is a single objective covering only the specifics of ingest, although the other objectives and procedures in the guide go well beyond those needed for ingest to the subsequent preservation treatment, access control, and rights management of objects. PAIMAS similarly describes a fairly restricted Submission Agreement, but includes

consideration of future financial, technical and management issues in the steps leading up to the Agreement. In fact, although both specifications profess their scope is the transfer of information, the transfer and ingest of SIPs can not realistically be considered outside of the broader context of a long-term archiving agreement.

An inter-repository service agreement, as envisioned by TIPR, must clarify the technical details of a specific act of transfer, but it must also explicitly address post-ingest preservation treatment, ongoing access controls, rights, and communications.

5. THE INTER-REPOSITORY SERVICE AGREEMENT

The last section explored the general applicability of PAIMAS and the nestor Guide to the case of repository to repository transfer. This section focuses specifically on the inter-repository service agreement as a variant of the Submission or ingest agreement. The TIPR approach was to define a relatively rigid transfer format for machine processing and rely on the inter-repository service agreement to provide context, meaning, and external stipulations.

5.1. Meaning of RXP elements

The RXP defines a standard place to put some critical pieces of information, but does not define code lists (controlled vocabulary) or semantics for the content. For example, the sending repository is identified in the *agent* element of the METS header in rxp.xml. The value used for identification must be negotiated between the parties and documented in the inter-repository service agreement. The receiving repository may need to predefine an agent record, add a mapping to a processing table, etc. This also applies to identification of the original producer and the original rights holder.

5.2. Transfer details

The RXP specification defines only a transfer format, and leaves details of the transfer protocol to be determined by the parties. In the TIPR project, test packages were bundled according to the BagIt specification and transmitted via HTTP, but they could equally as well have been zipped in native form and shipped on a portable drive. The inter-repository service agreement should document agreement on the transfer mechanism and serialization, and manifests used (if any). In addition, communication between repositories and the handling of transmission errors must be specified. Transfer requirements are well covered in PAIMAS and the nestor Guide.

5.3. Actions to be taken on ingest

Actions taken by the receiving repository after successful transfer are out of scope for TIPR and the RXP. Whether and how the receiving repository performs quarantine, validates packages and files, gives notification of rejection or successful ingest, and gives notification of anomalies and non-fatal errors all must be agreed upon and documented. Although much of this is covered in PAIMAS and the nestor Guide, both specifications stop at the point where the receiving repository has validated and accepted responsibility for the SIPs, which for some preservation repository systems may be far in advance of the creation and storage of a new AIP.

A complication in repository-to-repository transfer is the circumstance that in some cases notification should be made to the sending repository, and in other cases to the original owner of the content. Especially in the case of succession, the receiving repository may need to establish an ongoing relationship with the original owner(s).

5.4. Archiving policies and responsibilities of the receiving repository

Repository systems differ greatly in their internal data models and the type and amount of metadata they store. The TIPR project asserts that preservation repositories engaging in package exchange should be capable of understanding METS stucture and the semantics of PREMIS events. Beyond that, what metadata will be retained and what will be understood (in the sense that it will be maintained in a usable fashion) by the receiving repository is a matter for negotiation and documentation. Similarly preservation treatment, retention of versions, ongoing reporting, future dissemination and access are all appropriate for documentation in the inter-repository service agreement.

5.5. Rights and premissions

The TIPR RXP provides a place to record package-level rights. TIPR partners assumed repositories would use PREMIS rights statements, but any XML-encoded rights schema could be used if agreed-upon and included in the inter-repository service agreement. Rights governing individual files in the package, whether metadata or content, is not covered by the RXP specification and is entirely a mater of agreement among transfer partners.

5.6. Financial arrangements

Costs involved in the transfer project and ongoing custodial costs should both be documented along with the method for identifying and billing the appropriate party. In the case of succession, a likely scenario is that fixed costs of the transfer project are assumed by the terminating repository but ongoing custodial costs must be charged to the original producers.

5.7. Legal issues

The source repository can be assumed to have a standing legal agreement with its own Producers clarifying intellectual property rights, responsibility for copyright infringement, and liabilities and warranties governing damage to content, treatment of content, and provision of services. In the case of repository-to-repository transfer,

the legal relationship between the Producers and the original repository may carry over to the receiving repository but is more likely to require re-negotiation. Legal issues pertaining to the source repository must be considered separately from those pertaining to the original depositors, and documented in the inter-repository service agreement.

6. CONCLUSION

Two existing standards address the transfer of information from a producer (in OAIS terms) to a preservation repository. Although neither explicitly restrict their applicability to the original producer or content owner, neither consider the special case of a repository to repository transfer. The three use cases of interest to the TIPR project have different implications for the methodology of transfer and the circumstances considered. An inter-repository service agreement has much in common with a Submission (ingest) agreement, but must have a longer-term scope and take into account two producers, the producers of the original SIP and the proxy producer, the repository that creates the RXP for transfer.

7. REFERENCES

[1] Caplan, P., "Repository to Repository Transfer of Enriched Archival Information Packages", in *D-Lib Magazine*, v.14 no.11/12 (2008). Available at http://www.dlib.org/dlib/november08/caplan/11caplan.html

[2] Caplan, P., "Towards Interoperable Preservation Repositories (TIPR)", U.S. Workshop on Roadmap for Digital Preservation Interoperability Framework, 2010. Available at http://ddp.nist.gov/workshop/papers/03_08_Caplan_TIPR.pdf

[3] Caplan, P., Kehoe, W., Pawletko, J., "Towards Interoperable Preservation Repositories", *International Journal of Digital Curation,* v.5 no.1 (2010).

[4] Consultative Committee for Space Data Systems, Producer-Archive Interface Abstract Standard (CCSDS 651.0-B-1 Blue Book (2004). Available at http://public.ccsds.org/publications/archive/651x0b1.pdf

[5] nestor working group for long-term preservation standards 2009 – Into the Archive – a guide for the information transfer to a digital repository. Draft for public comment. Available at http://files.d-nb.de/nestor/materialien/nestor_mat_10_en.pdf