FP7-ICT-2007-3-231161

Deliverable D2.2.3

Strategy for Use of Preservation Metadata for within a Digital Library

with examples of use in audiovisual preservation

Nir Sherwinter (Ex Libris), Richard Wright (BBC)

28/12/2010

Document administrative table

Document Identifier	PP_WP2_D2.2.3_Strat_Pres_Metadata	Release	0
Filename	PP_WP2_D2.2.3_Strat_Pres_Metadata_R0_v1.00.pdf		
Workpackage and Task(s)	WP2 – Models and environments for long-term audiovisual content preservation WP2T2 – Metadata for long-term preservation		
Authors (company)	Nir Sherwinter, Ex Libris; Richard Wright, BBC		
Contributors (company)			
Internal Reviewers (company)	Kurt Majcen, Joanneum; Andy Stauder, Univ Innsbruck		
Date	20/01/2011		
Status	Release		
Type	Deliverable		
Deliverable Nature	Report		
Dissemination Level	Public		
Planned Deliv. Date	31/12/2010		
Actual Deliv. Date	28/12/2010		
Abstract	A description of how to actually use preservation metadata to maintain quality of audiovisual content in a *trusted digital repository* context, across basic preservation processes (multivalent, emulation, migration).		

DOCUMENT HISTORY

Version	Date	Reason for change	Status	Distribution
0.0	05/08/2010	First Draft	Outline	Confidential
0.1	14/12/2010	Ex Libris and BBC contributions	Working Draft	Confidential
0.2	20/12/2010	BBC completion – summary, conclusions; for internal review	Final Draft	Confidential
0.3	28/12/2010	Includes internal review results	Delivered	Confidential
1.00	20/01/2011	Finalised and published	Release	Public

Table of contents

Scope

PrestoPRIME is the European publicly-supported project that addresses **preservation of digital audiovisual content**, and **access to audiovisual content in digital libraries,** using **Europeana** as our demonstration platform.

This document concentrates on *preservation metadata.* Digital Preservation succeeds either through intent (well-designed formal systems), or by accident (somebody finds things decades later, frozen in a barrel in Alaska). Because digital technology moves so rapidly, there is little chance that a barrel full of old IT storage hardware would be of much use decades later. Digital Preservation therefore concentrates on well-designed formal systems (based on *digital library* and *trusted digital repository* concepts), and a part of any such system is information on what is to be preserved, and how. This is the domain of *preservation metadata,* a relatively new concept but fundamental to a formal approach.

The scope of this document is preservation metadata as applied to formal approaches to preservation of file-based audiovisual content within formal systems (digital libraries, trusted digital repositories, digital preservation systems).

We present examples of use of preservation metadata (in real systems) to maintain quality of audio-visual content, across basic preservation processes (multivalent, emulation, migration). This report also provides examples of how processes can be semantically modeled and documented across the audiovisual content lifecycle, including metadata processing.

Executive summary

The last decade has seen the audiovisual world recognise the importance of metadata, with significant standardisation (SMPTE, EBU, ITU, AES) and implementation work (asset management systems). We now recognise that in a file-based world where content is no longer on shelves, metadata is our only mechanism for actually keeping – or finding – anything.

So, how do we keep metadata – and how do we use metadata to keep content? This document reviews the standards and definitions for *preservation metadata* (Section 1), reviews leading digital preservation repository implementations (Section 2: New Zealand National Library; Rosetta) and reviews all the studies (that we could find) on formal digital preservation of audiovisual content (Section 3) and studies, implementations, projects and prototypes for digital preservation of audiovisual content (Section 4).

Finally, in Section 5 we set out the use of *preservation metadata* – and the preservation *of* metadata – in the work of the PrestoPRIME project.

We find three related concepts:

1. metadata about preservation – the PREMIS framework for describing what is allowed to happen to objects in a controlled environment: a trusted digital repository;
2. preservation of metadata – the procedure used in a trusted repository for ensuring that not only is content preserved, but also metadata is preserved;
3. use of metadata within a trusted repository – these repositories need to 'understand' their content, at least along certain dimensions, and that understanding is held in metadata.

The PrestoPRIME contribution in these areas is given in Section 5. In summary:

- PrestoPRIME implements metadata about preservation, following the PREMIS approach. PREMIS is basically a *data dictionary* setting out different specific kinds of *actions, permissions, events* and *agents* – and their *relations*.

- PrestoPRIME also implements a full system for preserving all the metadata that enters its own PrestoPRIME repositories (an extended Rosetta system, and the new P4 = PrestoPRIME Preservation Platform). The full architecture for holding metadata in P4 is described in *D5.2.1 Architecture Design.* The essentials are given below in Section 5.1.

- Finally, PrestoPRIME has developed a metadata mapping tool, so that the semantics of diverse kinds of metadata can be used in the operation of a trusted digital repository.

A digital preservation system needs to take certain *preservation actions.* If no actions were ever needed, there would be no need for such a system. Typical actions are:

- migrating content from one file type to another
- re-wrapping content into a new wrapper (to replace, for instance an obsolete wrapper).

Such actions are typically performed on a whole class of objects, so at the very least, a digital repository has to 'understand' metadata coming into the system that is used to define such classes.

As an example, all content with MPEG-2 coding of a particular profile might be affected, or all content with a particular aspect ratio, or requiring a particular decoder or a certain kind or category of player (or other rendering application). Unfortunately not all files coming into a system have identical labelling as to such technical and practical aspects – hence the need for metadata mapping, so that the repository can 'understand' at least those characteristics used to define the categories and classes involved in preservation actions.

1 Definitions and Standards

An understanding of *preservation metadata* begins with:

- where the concept arose, and why;
- the differences between *preservation metadata* and other classes of metadata

1.1 Definitions

There are various ways to divide or categorise the information that accompanies file-based content. The categories are needed because metadata is used for different purposes. We build catalogues of (mainly) descriptive information about the content of files. People don't generally come to a library to find books of a particular size and shape, or type of paper. Similarly we expect users of a digital library to base most searches on descriptive information of the content.

However once an item is found, the other kinds of information do become relevant. A book that can be circulated has a practical difference from a book that is for reference use only (within the library). Similarly, an audiovisual file that can be played on a user's computer is significantly different from one that cannot – and this difference is part of the technical metadata.

While there is no single classification system for 'kinds of metadata', the following are common types:

- **structural** – how digital content is *packaged*, eg use of the METS standard for a resource that consists of more than one file
- **descriptive** – resource identification and description
- **technical** – format, storage, location and technical requirements
- **preservation** – data on preservation actions and requirements: events, agents, rights and permissions, and their relationships
- **rights** – data about use of the resource
- **administrative** – information needed to maintain a resource, such as subscription information for a technical journal

The Harvard University Library lists standards for the first five[1], but also lists all six as 'key metadata standards'. More information on categories of metadata is also available from JISC[2], covering:

1. **Structural metadata.**
2. **Descriptive metadata.**
3. **Administrative metadata.**
4. **Technical metadata.**

The origin of *preservation metadata* was the recognition that file-based content had problems of standardisation and of obsolescence. Digital libraries started with documents (text), and arose from a situation in the 1990's where there were many kinds of text documents (many competing brands of word-processor software) which were not compatible, and which did not necessarily last for more than a few years. How can a digital library be made of files which have compatibility and obsolescence problems from the beginning?

One answer was to change the files in some way: convert them all to a common standard, and migrate them if their format became obsolete, or headed that way. Conversion and migration are preservation actions.

Changing a digital object in a digital repository is a large step, and could even be seen as a violation of a basic principle of archiving: keep the original. But the very digital technology that makes such a

[1] http://hul.harvard.edu/ois/digproj/metadata-standards.html
[2] http://www.jiscdigitalmedia.ac.uk/audio/advice/metadata-and-audio-resources/

repository possible was also forcing these changes, these actions, because of the compatibility and obsolescence issues – which are still very much with us in 2010.

The point of preservation metadata is to be as clear as possible about what kinds of actions can be taken (permissions), along with when (events) and by whom (agents).

Once people had arrived at understanding the need to be formal about actions taken on objects within a digital repository, they developed a standard for describing these actions. That standard is PREMIS[3], developed in 2002-2004. It was seen as such a major step that it won the prestigious Digital Preservation Award in 2005[4]. It is basically a data dictionary setting out different specific kinds of actions, permissions, events and agents – and their relations.

1.2 Standards

Really there is one major international standard for Preservation Metadata: PREMIS. However because the context is formal repositories and the preservation of their content, PREMIS is closely linked to concepts within the OAIS[5] framework for digital preservation, and the associated METS[6] standard for constructing the *information packages* required by OAIS.

The PREMIS standard was defined by an international working group, organised by the Library of Congress[7]. The standard was last updated in 2008[8]. While it can by no means claim to be widely adopted by people storing files (because only a minority use formal digital repositories, and a minority of them have formal digital preservation strategies), it is still the only standard for preservation metadata, in itself an achievement in an era where a major problem with metadata is that there are 'so many standards to choose from'.

[3] http://www.loc.gov/standards/premis/
[4] http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/premis-data-dictionary
[5] http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683
[6] http://www.loc.gov/standards/mets/
[7] Op cit http://www.loc.gov/standards/premis/
[8] Lavoie, B. (2008) PREMIS With a Fresh Coat of Paint: Highlights from the Revision of the PREMIS Data Dictionary for Preservation Metadata. D-Lib Magazine 14(5/6)
http://www.dlib.org/dlib/may08/lavoie/05lavoie.html

2 Preservation Metadata Strategy as Implemented by Ex Libris' Rosetta and the National Library of New Zealand

Definitions and standards acquire and release value through actual use. One of the organisations that led the definition of PREMIS was the National Library of New Zealand. Now they are one of the leading developers of a formal digital preservation system, putting the PREMIS standard into practice. The following sections describe this real-world implementation.

Digital repositories are meant to preserve digital objects, which can be complicated. They can consist of many related files, hence the need to build formal 'information packages' described by *structural metadata* following a standard (e.g. METS). Further, the files are complicated, being a mix of content (e.g. text or audiovisual signals represented numerically) and metadata about the files and the content. This metadata needs preservation, so one thing preservation metadata does is define processes for preservation of other metadata, as well as for the preservation of anything else in the *information package.* The discussion that follows refers to all sorts of metadata (technical, rights etc) and a preservation strategy for preserving such metadata – and then also refers to PREMIS and preservation metadata as formalisms for this preservation activity. The authors hope that the reader will not get distracted by this mix of levels:

1. metadata about the preservation process itself;
2. metadata actually being acted upon (preserved) within a preservation process.

The first level is about *preservation metadata* as defined and standardised by PREMIS. The second level is about preservation **of** metadata – any kind of metadata.

2.1 National Library of New Zealand Experience

Steve Knight in his article *"Preservation Metadata: National Library of New Zealand Experience"*[9] explained that the National Library of New Zealand, has been actively engaged in work on preservation metadata. This has involved the development of a preservation metadata schema, a more granular implementation-ready data model/XML schema, a software application for programmatically extracting preservation metadata, and finally a repository for storing the gathered preservation metadata.

The Library's work on metadata began in 2000 and was based on the taxonomy described in Anne Kenney and Oya Reiger's Moving Theory into Practice: Digital Imaging for Libraries and Archives (2000)[10]: resource discovery, structural, rights management and access control, technical and administrative. Initial work concentrated on metadata for resource discovery (National Library of New Zealand, 2000)[11] and described the core descriptive metadata standards to be used by the Library for resource discovery across all media and for all the Library's collections.

[9] LIBRARY TRENDS, Vol. 54, No. 1, Summer 2005 ("Digital Preservation: Finding Balance," edited by Deborah Woodyard-Robinson), pp. 91–110; online at http://www.thefreelibrary.com/Preservation+metadata%3A+National+Library+of+New+Zealand+experience.-a0140239962

[10] Kenney, A., & Reiger, O. (2000). Moving theory into practice: Digital imaging for libraries and archives. Mountain View, CA: Research Libraries Group.

[11] Metadata standards framework for National Library of New Zealand. http://www.natlib.govt.nz/files/4initiatives_metafw.pdf

The Library released the first version of a logical model for preservation metadata online in November 2002[12], with a revised version incorporating lessons learned since the original version being made available in June 2003[13].

The National Library of New Zealand attempted to minimize the degree of overlap with other metadata and focused on that metadata necessary for preservation, including the notion that the preservation metadata record itself is an integral part of the preservation process.

In 2007, the National Library of New Zealand engaged Ex Libris and started to develop and implement Ex Libris' digital preservation solution. All conclusions from the research done by the library so far were taken into account during the design process of Rosetta.

[12]Metadata standards framework--Preservation metadata.
http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf
[13]Metadata standards framework--Preservation metadata (revised).
http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf

2.2 Ex Libris' Rosetta

The first Rosetta implementation was for the National Library of New Zealand (NLNZ). In fact, since NLNZ signed a development partnership agreement with Ex Libris, the product was designed jointly by NLNZ and Ex Libris.

As part of the Rosetta product design, both the Ex Libris system analysts and the NLNZ staff analyzed the aspects that influence preservation metadata such as:

1. The different types of material that should be preserved
2. The origin of the materials (it can be a system which generates the files automatically like a web archiving tool, it can be a scanned book or a producer that creates a video file)
3. Which tools should be used by Rosetta in order to extract technical metadata
4. Which types of metadata should be regarded as Preservation Metadata

The main aspect that was analyzed and always was part of each and every design phase of the system is similar to how PREMIS defines Preservation Metadata as "the information a repository uses to support the digital preservation process."

In the following sections this document will present how Rosetta was designed to support all the different preservation metadata types in a secure and accessible fashion for use when performing a preservation action.

2.3 Rosetta Preservation Metadata

Rosetta Preservation metadata includes the following:

- **Provenance Metadata:** what is the origin of the digital object, what changes have been made to the object i.e. the documentation of the chain of events and actions (as well as related agents) that a specific object has undergone in the repository.

- **Technical Metadata:** a comprehensive set of technical information about the object. The information can be derived from the file itself using an extraction tool or by a user that provides information about which application is needed to render and use the digital object.

- **Rights management:** assertions of one or more rights or permissions pertaining to an object that can answer the question: which intellectual property rights must be observed?

- **Authenticity:** is the digital object authentic? Can we make sure that we are preserving an object that wasn't unintentionally damaged or changed? Any authenticity information is also part of the provenance metadata.

- **Security:** an object must be stored securely so that nobody can modify it inadvertently or maliciously. Supporting metadata should be kept in order to support this requirement. For example: the only way to edit an object is by using the system, if an object is changed directly on the disk some of the metadata won't reflect this change and the system will reports the discrepancies

- **Storage:** Metadata can support media management by recording the type and age of storage media and the dates that files were last refreshed.

2.4 Rosetta Implementation and Support

2.4.1 Provenance Metadata

A Rosetta component named *Provenance Events Manager* is in charge of logging all provenance activities and creating the Provenance Metadata. Before storing the objects in the permanent repository, another process collects the Provenance Metadata and stores it as part of the AIP.

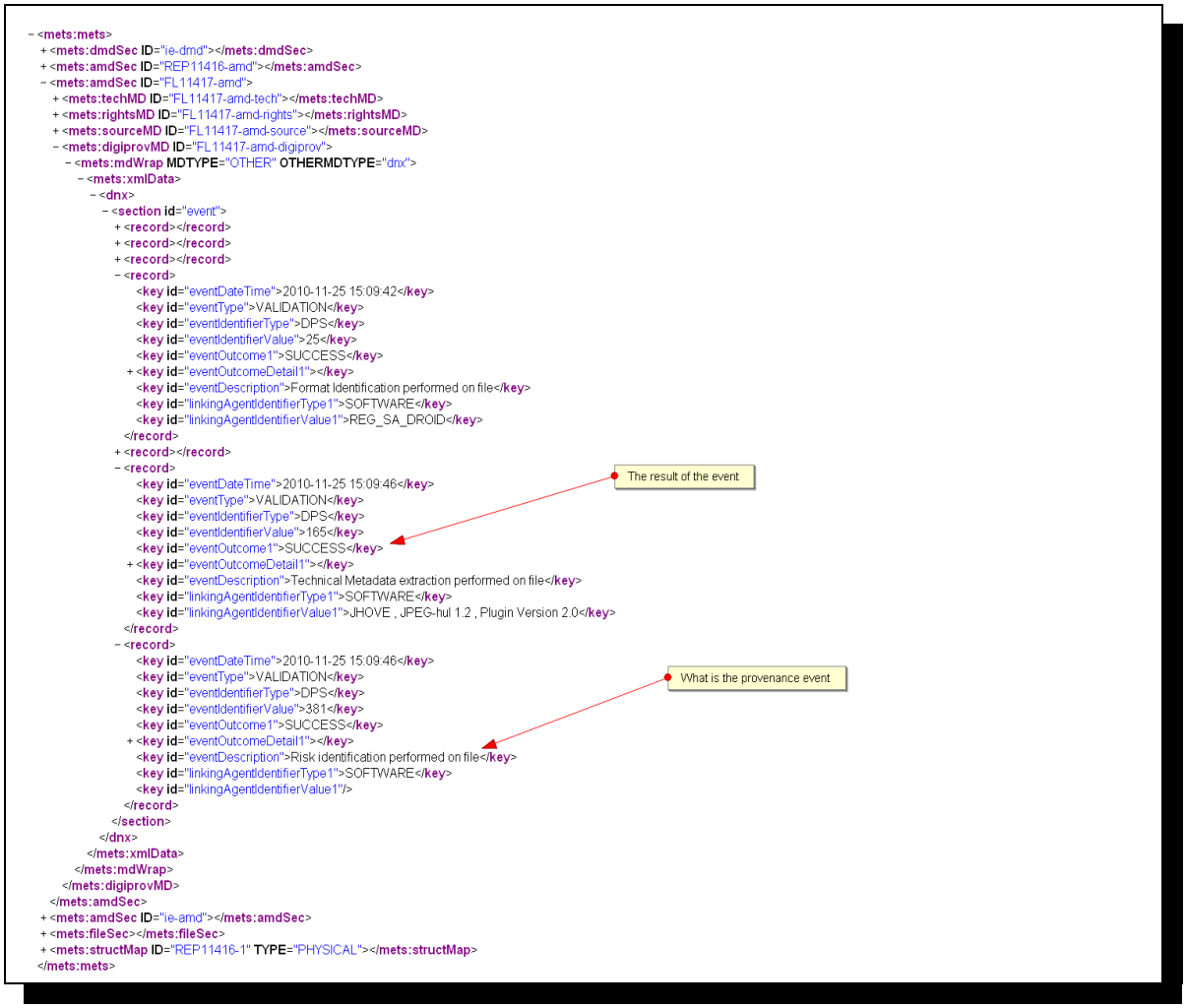
The following system activities trigger Provenance Events Manager processing:

- Changes to an object's metadata
- Changes to an object's structure
- Preservation action performed on the object

The provenance metadata can be used for reporting and providing the object's complete history.

For example, if it was announced that a major software company had released a tool for migration from MPEG-2 to MPEG-4 and as a result we would like to migrate all MPEG-2 to MPEG-4 files. Using the existing searching mechanism we can collect all of the MPEG-2 files, In addition - the provenance information can tell us about files that are currently not of the MPEG-2 type because they were migrated in a previous preservation action. We may consider rolling back the last migration action and migrate the files using the new migration tool.

An example of provenance information which is kept as part of the AIP:




```
- <mets:mets>
  + <mets:dmdSec ID="ie-dmd"></mets:dmdSec>
  + <mets:amdSec ID="REP11416-amd"></mets:amdSec>
  - <mets:amdSec ID="FL11417-amd">
    + <mets:techMD ID="FL11417-amd-tech"></mets:techMD>
    + <mets:rightsMD ID="FL11417-amd-rights"></mets:rightsMD>
    + <mets:sourceMD ID="FL11417-amd-source"></mets:sourceMD>
    - <mets:digiprovMD ID="FL11417-amd-digiprov">
      - <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="dnx">
        - <mets:xmlData>
          - <dnx>
            - <section id="event">
              + <record></record>
              + <record></record>
              + <record></record>
              - <record>
                  <key id="eventDateTime">2010-11-25 15:09:42</key>
                  <key id="eventType">VALIDATION</key>
                  <key id="eventIdentifierType">DPS</key>
                  <key id="eventIdentifierValue">25</key>
                  <key id="eventOutcome1">SUCCESS</key>
                + <key id="eventOutcomeDetail1"></key>
                  <key id="eventDescription">Format Identification performed on file</key>
                  <key id="linkingAgentIdentifierType1">SOFTWARE</key>
                  <key id="linkingAgentIdentifierValue1">REG_SA_DROID</key>
                </record>
              + <record></record>
              - <record>
                  <key id="eventDateTime">2010-11-25 15:09:46</key>
                  <key id="eventType">VALIDATION</key>
                  <key id="eventIdentifierType">DPS</key>
                  <key id="eventIdentifierValue">165</key>
                  <key id="eventOutcome1">SUCCESS</key>
                + <key id="eventOutcomeDetail1"></key>
                  <key id="eventDescription">Technical Metadata extraction performed on file</key>
                  <key id="linkingAgentIdentifierType1">SOFTWARE</key>
                  <key id="linkingAgentIdentifierValue1">JHOVE , JPEG-hul 1.2 , Plugin Version 2.0</key>
                </record>
              - <record>
                  <key id="eventDateTime">2010-11-25 15:09:46</key>
                  <key id="eventType">VALIDATION</key>
                  <key id="eventIdentifierType">DPS</key>
                  <key id="eventIdentifierValue">381</key>
                  <key id="eventOutcome1">SUCCESS</key>
                + <key id="eventOutcomeDetail1"></key>
                  <key id="eventDescription">Risk identification performed on file</key>
                  <key id="linkingAgentIdentifierType1">SOFTWARE</key>
                  <key id="linkingAgentIdentifierValue1"/>
                </record>
              </section>
            </dnx>
          </mets:xmlData>
        </mets:mdWrap>
      </mets:digiprovMD>
    </mets:amdSec>
  + <mets:amdSec ID="ie-amd"></mets:amdSec>
  + <mets:fileSec></mets:fileSec>
  + <mets:structMap ID="REP11416-1" TYPE="PHYSICAL"></mets:structMap>
</mets:mets>
```


In summary, provenance information reflects the complete history of the object, is part of the preservation metadata and is also part of the AIP. It is used by Rosetta for reporting as well as for analysis and decision making purposes (preservation actions wise).

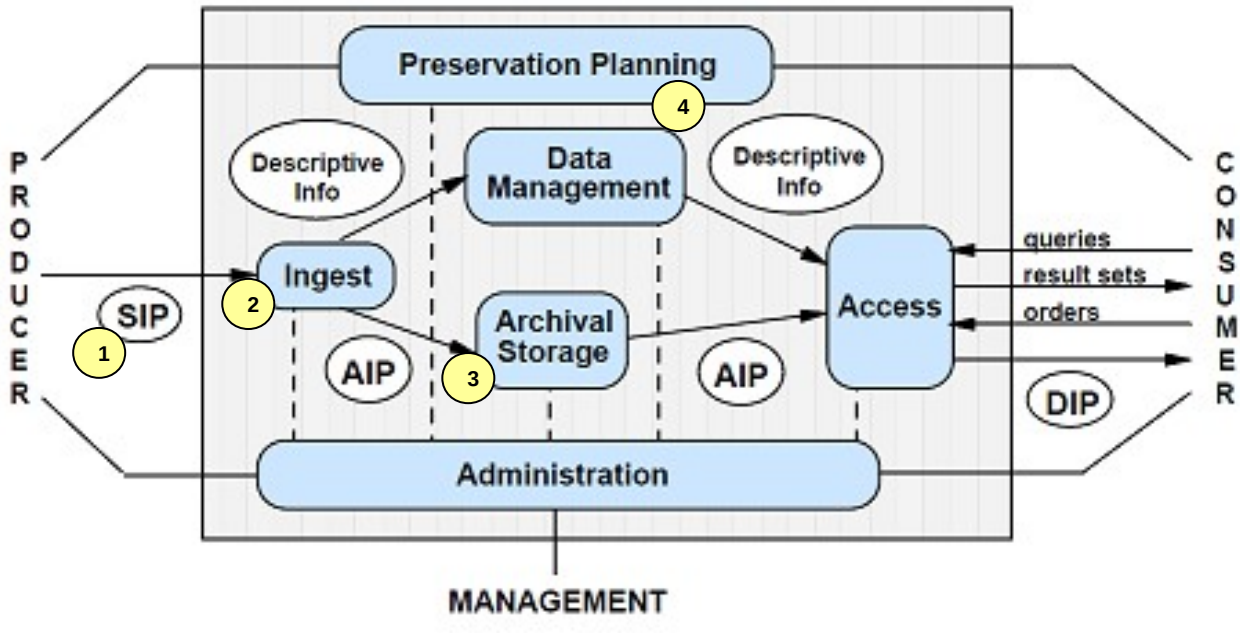
2.4.2 Technical Metadata

Technical metadata is one of the most important types of preservation metadata. OAIS, PREMIS and almost any reference model or standard defines Technical Metadata as something that should be part of any digital preservation system.

In order to preserve technical metadata of an object, the technical metadata should be made explicit first. Technical metadata can be created by the object creator and submitted into the system along with the object itself, or, can be extracted by the system during the ingest process.

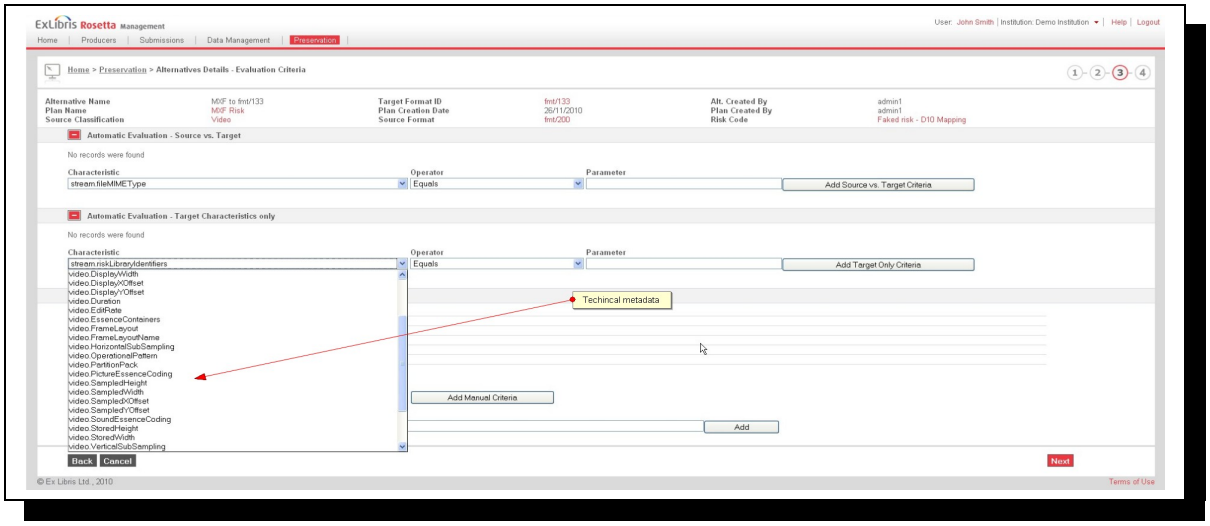
The technical metadata must be organized in a way that they can be useful as preservation metadata when preservation planning is required. For example, in order to compare two different video files, we need to get the resolution and other technical properties into a comparable form.

The following diagram illustrates the flow of technical metadata in a preservation system from the moment that the object is ingested till it is used as preservation metadata during preservation planning:



1. Technical metadata can be deposited by the producer
2. The system can extract metadata from the object using the technical metadata extraction tools such as MXF extractor, JHOVE, NLNZ etc.
3. The technical metadata are stored in the archival storage as part of the AIP
4. The preservation planning module:
    a. Searches the archival storage using the technical metadata
    b. Allows the preservation analyst to create the most effective preservation plans based on the technical metadata

The following screen shot is taken from the Rosetta preservation planning module. In this screen shot we can see how the preservation analyst can compare technical metadata of two objects: one is the source (the object with the risk) the other one is the target (the migrated object). The preservation analyst compares technical metadata in order to identify reduced quality or damage caused to the object during the migration process.



The technical metadata as shown above are kept in Rosetta as part of the IE in an XML structure. The following is an example of MXF technical metadata that are stored as part of the AIP:


```
- <section id="significantProperties">
  - <record>
      <key id="significantPropertiesType">video.VideoActiveLinesPerFrame</key>
      <key id="significantPropertiesValue">576</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.AudioChannelCount</key>
      <key id="significantPropertiesValue">8</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.SampledYOffset</key>
      <key id="significantPropertiesValue">0</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.DisplayXOffset</key>
      <key id="significantPropertiesValue">0</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.EditRate</key>
      <key id="significantPropertiesValue">25:1</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.Duration</key>
      <key id="significantPropertiesValue">1500</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.AudioSamplingRate</key>
      <key id="significantPropertiesValue">48000:1</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.DisplayWidth</key>
      <key id="significantPropertiesValue">720</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.SampledHeight</key>
      <key id="significantPropertiesValue">304</key>
    </record>
  - <record>
      <key id="significantPropertiesType">video.DisplayYOffset</key>
      <key id="significantPropertiesValue">16</key>
    </record>
```

In summary, technical metadata are one of the most important parts of the preservation metadata. They are used for identifying preservation risks, planning preservation actions and evaluating the results of migration processes. An AIP must contain all available technical metadata.

2.4.3 Rights metadata

There are two segments of rights metadata which come into consideration when designing a preservation metadata strategy:

1. Preserving the rights metadata i.e. embedding them as part of the AIP
2. Using the rights metadata in order to control access to the preserved object

Both should be supported by a preservation system. For the preservation of the rights segment, it is required that the rights metadata can be part of a deposited SIP and the preservation system will put it as part of the AIP. For the usage of the rights metadata in order to control access, the preservation system must have the ability to allow access and restrict/block access according to the definition in the rights metadata.

The second segment requires integration with a DRM system, which currently is not part of Rosetta.

Another aspect that should be taken into account is – do the rights metadata allow the holder of the object to perform a migration process? For example, the rights metadata explicitly state that copying the object is forbidden, but, a risk that was identified by the system shows that the object must be migrated… then – the preservation analyst must check the rights metadata and decide if he can perform this migration action or not.

The following diagram displays rights metadata embedded into an AIP (METS container):

```
- <mets:rightsMD ID="RMD1">
  - <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="METSRights">
    - <mets:xmlData>
      - <rts:RightsDeclarationMD RIGHTSCATEGORY="COPYRIGHTED">
        - <rts:RightsHolder>
            <rts:RightsHolderName>Fukami, Sueharu, 1947-, Japanese</rts:RightsHolderName>
          - <rts:RightsHolderContact>
            - <rts:RightsHolderContactAddress>
                Ruth and Sherman Lee Institute for Japanese Art, Hanford, CA, USA. See http://www.shermanleeinstitute.org.
              </rts:RightsHolderContactAddress>
            </rts:RightsHolderContact>
          </rts:RightsHolder>
        - <rts:Context CONTEXTCLASS="GENERAL PUBLIC">
          - <rts:Constraints>
            - <rts:ConstraintDescription>
                Transmission or reproduction of materials protected by copyright beyond that allowed by fair use requires the written permission of the
                owner. Responsibility for any use rests exclusively with the user.
              </rts:ConstraintDescription>
            </rts:Constraints>
          </rts:Context>
        </rts:RightsDeclarationMD>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:rightsMD>
```

2.4.4 Authenticity

Authenticity is part of the preservation metadata and should be kept as part of the AIP. In order to keep the object authentic by putting all authenticity information into the AIP, Rosetta includes two different functionalities:

1. The provenance and events mechanism that was described above in section 2.4.1
2. A mechanism that records information about the depositor, i.e. who (person or machine) sent the materials to the preservation system

The provenance mechanism was already described in this document. With regard to the second functionality, as part of the SIP processing the information about the depositor becomes part of the IE which later on becomes the AIP. Two types of preservation metadata for supporting authenticity are supported:

1. Authenticity information that is provided by the depositor as part of the SIP
2. Authenticity information that is recorded by the system during SIP processing

The following diagram displays an AIP which contains authenticity information that was logged by the system:

```
<mets:digiprovMD ID="ie-amd-digiprov">
 - <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="dnx">
   - <mets:xmlData>
     - <dnx>
       - <section id="producer">
         - <record>
             <key id="userName"/>
             <key id="address1"/>
             <key id="address2"/>
             <key id="address3"/>
             <key id="address4">Singapore</key>
             <key id="address5"/>
             <key id="defaultLanguage">en</key>
             <key id="emailAddress">Email@Email.com</key>
             <key id="firstName">NLB Internal</key>
             <key id="jobTitle"/>
             <key id="lastName">NLB Internal</key>
             <key id="middleName"/>
             <key id="telephone1">8888</key>
             <key id="telephone2"/>
             <key id="authorativeName">NLB Internal</key>
             <key id="producerId">40283</key>
             <key id="userIdAppId">40279</key>
             <key id="webSiteUrl"/>
             <key id="zip"/>
           </record>
         </section>
       - <section id="producerAgent">
         - <record>
             <key id="firstName">Yaniv</key>
             <key id="lastName">Cohen</key>
             <key id="middleName"/>
           </record>
         </section>
       </dnx>
     </mets:xmlData>
   </mets:mdWrap>
</mets:digiprovMD>
```

Another aspect of authenticity information is keeping it for future use. It can occur that an object is at risk but no good migration is found. In this case, the preservation analyst can try to contact the depositor and verify if other copies of different types exists.

2.4.5 Security

This type of preservation metadata is kept as part of the AIP and in the system. Using this method, the system can always make sure that an object is kept safe and unharmed. For example, the modification date is kept as part of the AIP and in the system, when the dates are different the system will notify the discrepancies.

Additionally, the system security mechanism that manages user privileges is also regarded as preservation metadata so far as it is information that allows securing the objects and making sure they are safe for the long term.

2.4.6 Storage

Metadata can support media management by recording the type and age of storage media and the dates when files were refreshed for the last time. This kind of information is not part of the AIP but is still regarded as preservation metadata since it is used for making sure that objects are not damaged due to lack of information about the storage.

3 *Preservation Metadata* in Digital Libraries/Repositories

We have gone into the National Library of New Zealand case in detail, in order to describe a real-world system and give substance to the notion of preservation metadata. We have full information because Ex Libris, the system developers, is a PrestoPRIME partner. However there are other examples of major digital repositories with formal implementations of preservation metadata.

- **CDL = California Digital Library**[14]; from their website:
  "The CDL was founded by the University of California in 1997 to take advantage of emerging technologies that were transforming the way digital information was being published and accessed. Since then, in collaboration with the UC libraries and other partners, we assembled one of the world's largest digital research libraries and changed the ways that faculty, students, and researchers discover and access information."

  The technical architecture of the preservation repository of the CDL[15] dates from 2000-2003, with the last report published February 2004. Hence it predates PREMIS (to some extent), and indeed the technical architecture document makes no mention of PREMIS.

  However the concepts are there:
  1. There are *agents* that can act on a version of a digital object, according to a set of defined *behaviours.*
  2. There is a separation between *metadata database* and *metadata storage,* fundamental to the ability to preserve metadata.

- **SDR = Stanford Digital Repository**[16]. This is a repository serving a single university, but one of the world's major technical universities. The repository has an architecture that has recently undergone a major revision . The initial system was made operational in 2006, and a full revision was implemented in 2009-2010[17]. Again, there is no specific mention of PREMIS. However SDR recognised the need for preservation of metadata as a critical function of digital preservation, to the extent that they introduced the open-source repository system Fedora[18] specifically as a metadata repository within the overall digital object repository.

- **LC = Library of Congress**. The Library of Congress runs the major digital preservation project anywhere, the National Digital Information Infrastructure and Preservation Program[19] – a $100 million group of several dozen projects, all on digital preservation. For the LC itself, the University of California at San Diego (UCSD) performed a study on preservation repository requirements[20]. UCSD already has significant experience with holding image data, including an NDIIPP study on preservation requirements of video. PREMIS was specifically included in the report, as one of two tools (the other one is METS) needed to transfer digital files from the LC to the UCSD supercomputer centre, for long-term preservation.

- **NARA = National Archives and Records Administration**[21], the archives of the US government. They have their own very large digital repository project[22], but they also developed (with the Research Libraries Group) a checklist[23] for the evaluation of digital repositories. One important area of the checklist is *B3. Preservation planning, migration, &*

[14] http://www.cdlib.org/
[15] http://www.cdlib.org/services/uc3/docs/DPRDesign_Final_Dev.pdf
[16] http://lib.stanford.edu/sdr
[17] http://www.dlib.org/dlib/september10/cramer/09cramer.html
[18] http://www.fedora-commons.org/
[19] http://www.digitalpreservation.gov/library/
[20] http://www.loc.gov/standards/premis/PREMIS_LCProfile.ppt
[21] http://www.archives.gov/
[22] http://www.archives.gov/era/
[23] http://www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file2433.html

other strategies. While PREMIS is specifically mentioned only once, the whole section depends upon data about the preservation process: preservation metadata.

- **TNA = The National Archives**. These are the archives of the UK government, which have their own very large responsibility for preserving digital content. Their *Active Preservation Framework* is a digital repository and preservation system[24]. The data model is said to pre-date PREMIS, but in the article sited they describe reviewing their model against PREMIS, as follows:

  > *"... TNA has reviewed and mapped the scheme against PREMIS: this has been useful in identifying much common ground, as well as some important differences, such as the way PREMIS addresses* significant properties*. The record entity describes the conceptual electronic records being managed, and corresponds to the PREMIS information object. Manifestations describe the particular technical instantiations of that conceptual record, and are equivalent to the PREMIS concept of representations. A given manifestation is composed of one or more files and bitstreams, which are defined in accordance with PREMIS. Both records and files may have properties associated with them, the uses of which are elaborated in Preservation Planning ..."*

- **BL = British Library**. This institution is the national library for the UK, and is one of the world's major national libraries, with a large amount of digital preservation activities – such as running the recent Planets project[25] (on digital preservation planning). As part of Planets, the BL presented various reports on their own digital preservation approaches and technology, including " PREMIS Implementations at the British Library[26]" – a report specifically on preservation metadata and BL's implementation of the PREMIS standard.
  PREMIS plays a central role in BL preservation work, and is described as a "blue-print for capturing preservation metadata for archival information packages"[27].

[24] http://www.ijdc.net/index.php/ijdc/article/viewFile/37/10
[25] http://www.planets-project.eu
[26] http://www.planets-project.eu/docs/presentations/Dappert_Rome2009.pdf
[27] Op cit, slide 3

4 Preservation Metadata and Audiovisual Content

The motivation for PrestoPRIME was to connect the preservation needs of audiovisual content with the preservation technology from the digital library world. The problem is that digital libraries were designed around text, or text and images (particularly images of text, from book-scanning projects). The result is that there are few examples of formal digital repositories and related technology (OAIS, METS, PREMIS) working on audiovisual content.

However there are three systems that use mainstream digital preservation technology, and have as their *main* goal the preservation of audiovisual content:

- **PDPT=Preserving Digital Public Television**, an NDIIPP-sponsored project between Channel 13 (the New York public service broadcaster) and New York University; the project used an OAIS architecture, METS structures for information packages, and PREMIS preservation metadata.
- **Memories**, an EC project based in Brussels, and focussed primarily on audio content; in addition to an OAIS architecture (complete with rich metadata), the project also implemented a semantic approach for gathering metadata within a well-defined context (ontology), so that associations could be automatically created.
- **Archipel**, a prototype at the Multimedia Lab of the University of Ghent for a digital archive for Flanders, specifically addressing audiovisual content.

4.1 PDPT=Preserving Digital Public Television

This project developed a complete digital repository and preservation system for audiovisual content, designed for a television broadcast company. All those factors make this the most relevant (to PrestoPRIME) of all existing work in digital preservation.

The project is fully-described in its final report, Preserving Digital Public Television[28], published in July 2010.

The work ran over a seven-year span, with the main effort under NDIIPP (the Library of Congress National Digital Information Infrastructure and Preservation Program) sponsorship.

Preservation metadata was explicitly created during the production of OAIS-standard Archive Information Packages (AIP). An AIP is the basic storage unit of the OAIS approach, and preservation metadata is created to fulfil several functions:

- preservation of the metadata within the AIP;
- PREMIS metadata to control preservation of the AIP itself (metadata controlling and documenting preservation actions)

The PrestoPRIME project has tried to learn as much as possible from PDPTV about the successful and efficient production of information packages and associated metadata (preservation or otherwise), which is now (end of 2010) being implemented in the PrestoPRIME preservation system.

Creation of PREMIS preservation metadata was not seen as a major difficulty in the PDPTV work, though there are a lot of standards with their associated complexity, as described in the entertaining PDPTV powerpoint “PBCore, METS, PREMIS, MODS, METSRights...oh my![29]” .

[28] http://www.thirteen.org/ptvdigitalarchive/uncategorized/final-report-preserving-digital-public-television/124/
[29] http://www.google.co.uk/url?sa=t&source=web&cd=1&ved=0CBgQFjAA&url=http%3A%2F%2Fwww.slideshare.net%2Fkvanmalssen%2Fpbcore-mets-premis-mods-metsrightsoh-my&ei=m34GTcrgE9-qhAez4uDtBw&usg=AFQjCNG1as6oWwsK7fGD4EIlE-JpZpgiyA&sig2=3VSdKwAQ4whMZsjE6Tmyhw

A key metadata problem faced by PDPTV was getting metadata from audiovisual files into a form that allowed the metadata to be kept separately in the repository, making preservation independent of the idiosyncrasies of particular audiovisual formats. This was solved for the purposes of PDPTV by making sure the system had metadata extraction tools that worked on the specific file types that were the 'house standards' of Channel 13.

PrestoPRIME addresses the preservation of a larger range of metadata through the use of a metadata mapping tool (developed by the project) that can process several standard kinds of audiovisual metadata (Dublin Core, EBU-Core, PB-Core, P/META and MPEG-7).

4.2 Memories

Memories[30] was an EC project based in Brussels, and focussed primarily on audio content. However the coordinator was the commercial company Memnon, which has integrated key Memories technology into its own preservation processes.

Memories had two interests:

- digital preservation following recognised standards
- powerful search-and-retrieval using semantic associations to metadata terms (tags)

The implementation of OAIS was quite standard, and concentrated on audio where there are not so many file formats and where there is considerable agreement on the Broadcast version[31] of the well-known WAV format.

The real complexities of Memories were around the effort to implement meaningful associations (RDF=Resource Description Framework[32] links) for metadata terms, and do this in a semi-automated way (using an ontology that was customised for the 'profile' of the operation/operator in question). This work was (in the author's opinion) very advanced and surprisingly successful – but a discussion of the work is outside the scope of this document, as this part of the Memories project was about creation of rich metadata, not about either preservation metadata nor the preservation of metadata.

4.3 Archipel

The only other active project using OAIS and preservation metadata and specifically addressing audio-visual content is *Archipel*[33], from the Multimedia Lab of the University of Ghent. The project is building a prototype general digital archive for the Flanders region of Belgium (thus not quite a national, but rather a 'near-national' archive). Because of the amount of audiovisual content now being generated and shared (particularly in a web context), and because of the high heritage and historical interest in such content, the project specifically addresses audiovisual and multimedia issues.

Like Memories, Archipel also combines metadata and semantics in a new way. Archipel has updated the concepts underlying PREMIS, and recast them according to the principles and practices of semantic web technology. So whereas Memories looked at the general problem of making tags compatible with RDF, Archipel has specifically focussed on PREMIS metadata itself.

[30] http://www.memories-project.eu/
[31] http://www.cube-tec.com/knowhow/bwf.html
[32] http://www.w3.org/RDF/
[33] http://www.ibbt.be/en/projects/overview-projects/p/detail/archipel

Archipel has created an ontology to add a semantic dimension to PREMIS relationships[34]. The purpose of the work is, ultimately, to literally 'add meaning' to the use of preservation metadata. This work should allow preservation actions to potentially 'make sense' in ways that are not possible for the current PREMIS approach and for current implementations (e.g. Rosetta).

PrestoPRIME will look carefully at the progress of Archipel, and if the 'preservation ontology' will either add new capabilities or simplify existing ones, we will evaluate adding the approach to the PrestoPRIME Preservation Platform (P4)[35].

4.4 Other relevant work

Here is a short list of other OAIS-based preservation work specifically addressing audiovisual content. Most are research projects that have now been closed, or new projects that have not yet achieved significant results.

- **VidArch**[36] "This project focused on developing a preservation framework for digital video context by applying it to two important digital video collections: the complete series of NASA broadcast educational videos and the complete set of juried ACM SIGCHI videos presented at annual conferences from 1983 to the present.

  The project addressed the important context aspect of digital preservation on both theoretical and practical fronts, which should improve archival decision-making and finding-aid creation and suggest ways to leverage technology further to make them more efficient and effective."
- **MediaMap**[37] This is a project that is in some ways picking up where Memories left off, and so is of continuing interest to PrestoPRIME.
- **Caspar**[38] This project looked at video art and other forms of cultural heritage materials, and their digital preservation using OAIS. PrestoPRIME has studied the results of Caspar.
- **EDCine**[39] This project was specifically about digital cinema. It made a 'light' implementation of OAIS, in which a single 'wrapper' file (MXF file) was used as an OAIS implementation package. This avoided virtually all the complexities of OAIS, leaving open whether or not it achieved any of the goals of OAIS. PrestoPRIME had a joint meeting with EDCine in London in May 2009, and has studied the EDCine results in detail.

[34] http://biblio.ugent.be/record/1081007 paper at AMIA-IASA, November 2010
[35] http://www.prestoprime.org/docs/training/P4-Allasia-Vienna2009-10.pdf
[36] http://www.ils.unc.edu/vidarch/
[37] http://www.mediamapproject.org/project_situation.html
[38] http://www.casparpreserves.eu
[39] http://ec.europa.eu/avpolicy/docs/reg/cinema/june09/edcine.pdf ; http://www.edcine.org/

5 PrestoPRIME Use of Preservation Metadata

PrestoPRIME is demonstrating formal approaches to digital audiovisual preservation in two ways:

1. additions to the approach of Ex Libris and their Rosetta product, to support time-based content
2. creation within the project of P4 = the PrestoPRIME Preservation Platform

The definition and use of preservation metadata in Rosetta has been fully described above, in Section 2 “Preservation Metadata Strategy as Implemented by Ex Libris Rosetta and the National Library of New Zealand”.

The use of Preservation Metadata, and the formal strategy for preservation of metadata, are described in the P4 Technical Architecture documentation, principally:

5.1 Submission Information Package (SIP)

The metadata that is implemented in the PrestoPRIME SIP is defined in Tables 5 and 6 of D5.2.1 Architecture Design. Here is that information:

Information Requirement	**Level**	**Can be Updated**	**Comments and Examples**
Identification metadata	mandatory	yes/no	Information for identifying the Editorial Entity - updatable: title, credits, etc. - non-updatable: actual Editorial Entity identifiers
Descriptive metadata	optional	Yes	
Technical metadata	Recommended		If present implies verification on ingestion, which would permit the detection of discrepancies with the provided information. If not present the technical metadata will be extracted by direct inspection of AV material and thus no discrepancy check is possible. This information is required to interpret the content. It can be duplicated in the content wrapper.
Rights metadata	Recommended	Yes	The digital rights metadata is provided by one OWL file. The definition of the rights metadata is under the responsibility of WP4T4
Ingest and Preservation Options and Service Terms	mandatory	yes	E.g. - permission/request to create a browsing quality copy - permission/request to create access copy - permission to change the file format - reference to a specific defined and agreed policy - preservation and delivery SLA
Provenance	mandatory, but history previous to ingestion: optional	yes	Some initial provenance information must be supplied by the Producer, such as: - which Producer and which ingest flow the content comes from and when it was ingested mandatory) - previous history (optional) The Archive will update the history section when a relevant event happens to the file within the Archive itself.

Update and access permissions	mandatory	yes	- group of users with update permission (content/rights/metadata/policy) - group of users with permission to search & browse (content/rights/metadata/policy) - group of users with access to delivery (quality level) User profiles need to be created

The “Ingest and Preservation Options and Service Terms” row in the above table is the preservation metadata. In particular, “permission to change the file format” is a fundamental preservation issue that PREMIS is designed to handle because PREMIS gives a formal structure for defining who has the permission, and even for defining what kind of changes are allowed. For now, the P4 development hasn’t reached that level of detail, but PrestoPRIME is aware of the suitability of PREMIS for preservation metadata in the SIP.

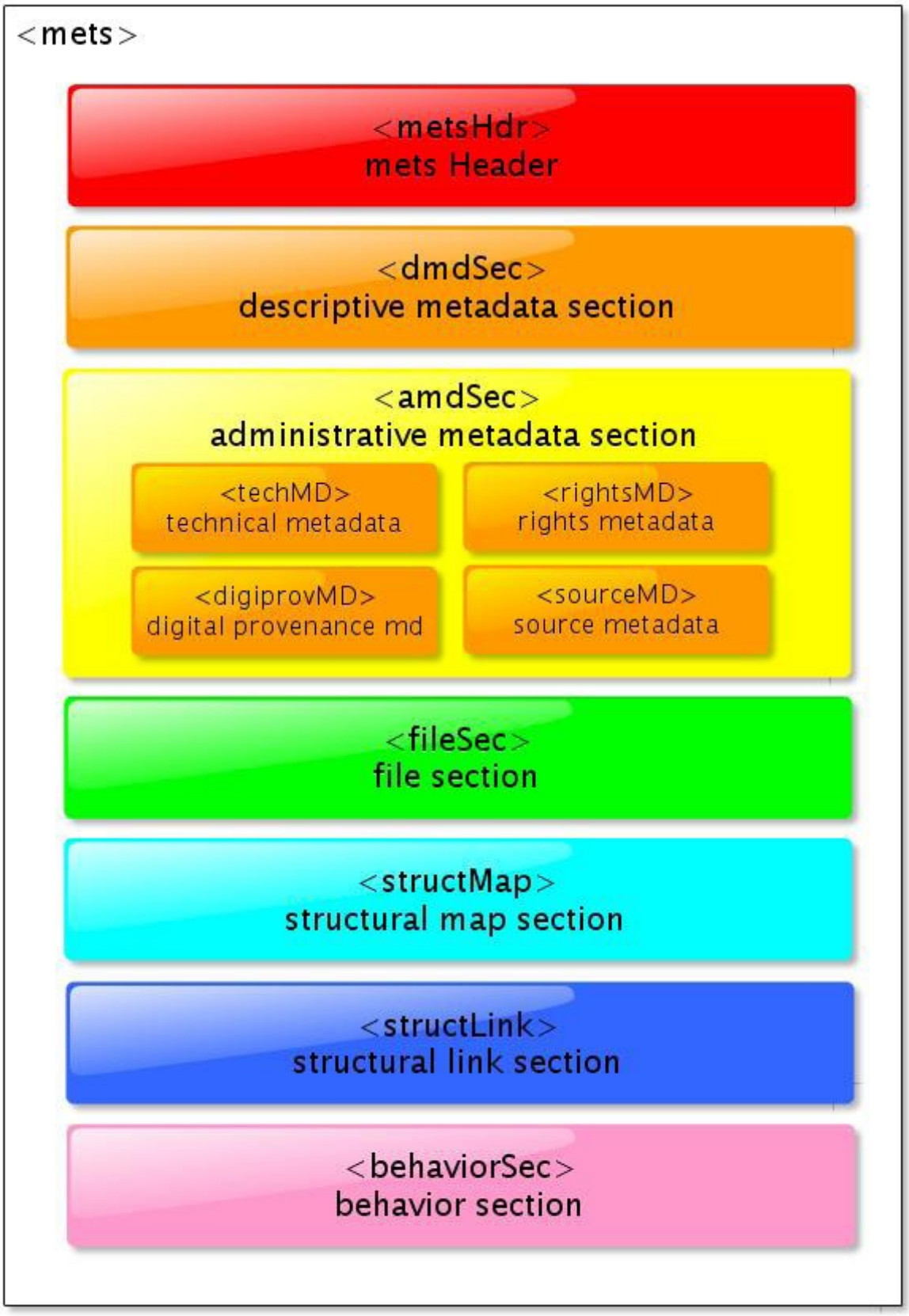
PrestoPRIME is using a METS wrapper to hold the incoming content and metadata as a SIP. METS has a definition of sections that, for simplicity, has two categories of metadata: descriptive and administrative. In this case, administrative metadata is used in a very general sense, basically meaning ‘everything besides descriptive metadata’.

In the P4 architecture (as shown in the following diagram, Fig 18 in D5.2.1), there are four kinds of administrative metadata:

- technical
- rights
- digital provenance
- “source metadata”

Two questions arise:

- What is “source metadata”? – a straight copy of metadata found in the source. Such metadata can also be acted upon by a metadata extraction tool, and prepared for active use in the digital preservation system – and so end up elsewhere in the SIP, but this section allows a simple way to capture all the source metadata.
- What happened to the *preservation metadata*, the “Ingest and Preservation Options and Service Terms”? – From D5.2.1, p 65: “The <techMD> is the section where the Ingest and Preservation Options and Service Terms are to be included. This section will probably make use of PREMIS (or DNX as a subset of PREMIS) ...”


<mets>
<metsHdr>
mets Header
<dmdSec>
descriptive metadata section
<amdSec>
administrative metadata section
<techMD>
technical metadata
<rightsMD>
rights metadata
<digiprovMD>
digital provenance md
<sourceMD>
source metadata
<fileSec>
file section
<structMap>
structural map section
<structLink>
structural link section
<behaviorSec>
behavior section

5.2 The preservation of metadata in PrestoPRIME P4

Throughout this document there have been two related issues: preserving metadata, and a particular kind of metadata that defines preservation processes: preservation metadata. The preceding section (5.1) gave a brief account of how preservation metadata is put into the SIP, and so carried into the P4 system.

The PrestoPRIME strategy for preserving metadata has three parts:

- creation of the SIP
- creation of the AIP
- operation of the P4

5.2.1 Creation of the SIP

The P4 will make use of different kinds of metadata, so the first step is the extraction of metadata from files that come in through an ingest process. Successfully extracted metadata can then be split into categories (descriptive, technical, rights, provenance) and placed into the appropriate part of the SIP, as shown in the above diagram.

For the P4 system to make use of metadata, there may need to be a mapping stage so that diverse kinds of metadata can be converted to a 'common language'. PrestoPRIME has a tool supporting standard kinds of audiovisual metadata, as already mentioned. The advantage of mapping is that it makes terminology usable (in P4). The disadvantage is that a mapping process changes the original metadata content, which violates the 'keep the original' archiving principle. For this reason, the un-mapped *source metadata* can also be carried straight into the SIP.

Source metadata that is not XML formatted creates a problem, and there are further problems with *dark metadata*[40] that may not even be interpretable by the ingest process. A brute force solution to raw preservation of that metadata is through keeping an exact copy of all such source metadata. This satisfies the 'letter of the law' in that nothing is lost, but this solution does not advance the preservation of such metadata – metadata that is not interpreted in any way is metadata that cannot be acted upon with any sophistication. The bits can be preserved, but the meaning will remain as dark as the *dark metadata* input.

5.2.2 Creation of the AIP

Within P4, the stored object is the AIP. The construction, storage and management of AIPs are the heart of the P4 system. The architecture and functionality of P4 is fully defined in *Section 5.2 The Preservation Platform Design* of deliverable D5.2.1 Architecture Design.

While D5.2.1 gives an explicit definition of a SIP, the full description of the AIP is subject to further work in PrestoPRIME, and will be documented in deliverable D5.2.2[41].

5.2.3 Operation of the P4

The use of *preservation metadata* is within the Preservation Planning component (D5.2.1, Fig 23). The operation of the Preservation Planning component is described in Section 5.2.3 of D5.2.1. The Preservation Planning component is a key part of the OAIS approach to digital preservation, and an explanation of that component is beyond the scope of this document.

We have described how preservation metadata comes into P4 and gets used, and we have also described the general strategy for preservation of all metadata. Further description of use of preservation metadata is contained in PREMIS and OAIS documentation, and further description of the P4 operation is contained in PrestoPRIME D5.2.1 (delivered) and D5.2.2 (forthcoming).

[40] Dark Metadata is the term given to metadata that is unknown by an application. http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies-20070724/

[41] PrestoPRIME Project. Deliverable 5.2.2:Prototype of open PrestoPRIME reference implementation, expected 31/12/2011.

Conclusions

We have considered three related concepts:

1. metadata about preservation – the PREMIS framework for describing what is allowed to happen to objects in a controlled environment: a trusted digital repository;
2. preservation of metadata – the procedure used in a trusted repository for ensuring that not only is content preserved, but also metadata is preserved;
3. use of metadata within a trusted repository – these repositories need to understand their content, at least along certain dimensions, and that understanding is held in metadata.

The PrestoPRIME contribution in these areas was discussed in Section 5. In summary:

- PrestoPRIME does implement metadata about preservation, following the PREMIS approach

- PrestoPRIME also implements a full system for preserving all the metadata that enters its own PrestoPRIME repositories (an extended Rosetta system, and the new P4 = PrestoPRIME Preservation Platform).

The full architecture for holding metadata in P4 is described in *D5.2.1 Architecture Design.* The essentials were given above in Section 5.1.

- Finally, PrestoPRIME has developed a metadata mapping tool, so that the semantics of diverse kinds of metadata can be used in the operation of a trusted digital repository.

Glossary

Term	Definition
AES	Audio Engineering Society
AIP	Archive Information Package
DIP	Delivery Information Package
DRM	Digital Rights Management
EBU	European Broadcasting Union
Information Package	The basic *object* in OAIS; there are three types: SIP, AIP, DIP = Submission (input), Archive (storage), Delivery (output)
IT	Information Technology
ITU	International Telecommunications Union
JHOVE	JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/ - a web service for identification, validation and characterisation of file formats
METS	Metadata Coding and Transmission Standard – one way to build OAIS information packages http://www.loc.gov/standards/mets/
MPEG	Moving Picture Expert Group – the body behind an important range of audiovisual encoding standards http://www.mpeg.org/
MXF	Media Exchange Format – a non-proprietary SMPTE standardised wrapper format for audiovisual content, used in broadcasting and digital cinema and related professional contexts
NLNZ	National Library of New Zealand
OAIS	Open Archival Information System – the ISO standard for a digital preservation system http://en.wikipedia.org/wiki/Open_Archival_Information_System http://public.ccsds.org/publications/archive/650x0b1.pdf
P4	PrestoPRIME Preservation Platform
PREMIS	***PREMIS (Pre****servation* ***M****etadata:****I****mplementation* ***S****trategies)* A framework for preservation metadata http://www.loc.gov/standards/premis/
RDF	Resource Discovery Format
Rosetta	The digital preservation product of Ex Libris
SIP	Submission Information Package
SMPTE	Society of Motion Picture and Television Engineers

References

PREMIS and its foundations:
http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata

http://www.ukoln.ac.uk/metadata/cedars/papers/aiw02 *(Cedars)*

http://www.nla.gov.au/padi/topics/32.html *(PADI)*

http://www.oclc.org/research/activities/past/orprojects/pmwg/presmeta_wp.pdf *(this is one of the foundation papers from 2001, before PREMIS)*

http://www.ariadne.ac.uk/issue22/metadata/ *Metadata for digital preservation: an update (this is Michael Day in Dec 1999 !)*

http://www.dpconline.org/advice/technology-watch-reports.html *DPC* ***Technology Watch Report - Preservation Metadata (2005; Lavoie and Gartner)***