SCAPE Digital Object Model

Authors

Person	Role	Partner	Contribution
Matthias Hahn		FIZ	
Frank Asseg		FIZ	
Nir Sherwinter		ExLibris	
Rui Castro		KEEPS	

Distribution

Person	Role	Partner
SCAPE ALL		All Partners

Revision History

Version	Status	Author	Date	Changes
0.1		Matthias Hahn	2012-03-14	
0.2		Matthias Hahn	2012-03-26	
0.3		Nir Sherwinter	2012-03-27	
0.4		Frank Asseg	2012-03-28	
0.5		Rui Castro	2012-03-29	
0.6	Final	Matthias Hahn	2012-05-07	
0.7	Post-final	Matthias Hahn	2012-08-15	added requirement for an identifier to be a string in sec 4.4


SCAPE

Table of Contents

1 Introduction

To be able to implement repository services like the Connector API and the Loader Application we need to agree on a Digital Object Model within the SCAPE project. Its obvious that each repository already provides a Digital Object Model but the diversity hinders the SCAPE platform to integrate all the partner respositories. The lack of a Digital Object Model has been put on the SCAPE risk register on Sharepoint[1]. This Document tries to resolve this issue.

In order to use the same terminology throughout the document we give a short introduction into the well-known standards like OAIS, METS and PREMIS used in the long term preservation world. Some questions related to our domain specific requirements are discussed.

On the most abstract level the reference model for archives, the OAIS model will be dicussed in brief. The METS standard describes a XML container for metadata structure of digital objects. METS is widely used for interoperability between repositories and service components. The PREMIS standard describes a semantic model for preservation metadata, and is widely used in long term preservation. Using METS and PREMIS together will be discussed briefly since there are some issue one has to be aware of.

Some of the repositories of the SCAPE members already use METS and PREMIS, but the mere employment of these standards does not guarantee interchangeability in between digital repositories. There might for example be significant differences in between two METS documents as they are used by two different repositories. We will describe the current existing data models of the repositories in this document and develop a possible model every repository holder may subscribe to.

The outline of the document is as follows: after a short introduction into the Open Archival System (OAIS), METS and PREMIS we are going to discuss the current digital object model of each repository of SCAPE. The next section will describe the DigitalObject Model that will be used within the SCAPE project.

2 Theoretical Background

This chapter deals with some basic understanding of wide spread standards and de-facto standards used in long term preservation projects. We start with most abstract description of a repository for long term preservation named Open Archival System. The purpose is to introduce the reader into the naming convention that we will use in this document. An introduction into METS and PREMIS and the interplay of both standards is given as well in this chapter. Understanding METS and PREMIS is crucial to understand the SCAPE Digital Object Model, because it builds on these standards.

[1] https://portal.ait.ac.at/sites/Scape/Management/Lists/SCAPE%20Risk%20Register/AllItems.aspx

2.1 OAIS

OAIS is the acronym for an Open Archival System and describes on an abstract level the requirements an archival system for long term preservation has to fulfill. The following six functional areas are described by the reference model:

1. Ingest
2. Archival Storage
3. Data Management
4. Administration
5. Preservation Planning
6. Access

The key terms for this document are SIP (Submission Information Package), AIP (Archival Information Package) and DIP (Dissemination Information Package).

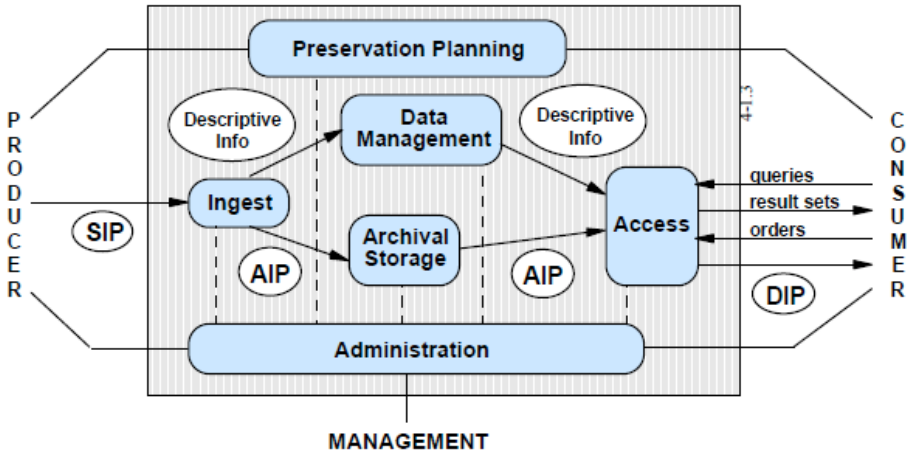


Figure 2.1-1: Illustration of the OAIS functional entities

The current SIP definitions of existing SCAPE repositories have significant differences: In RODA for example a SIP is a compressed ZIP file that contains a METS envelope, in Rosetta a SIP may contain several IEs (Intellectual Entities). An AIP contains technical metadata and metadata important for long term archiving.

2.2 METS

The METS specification (in XML format) conveys the metadata to manage digital objects within a repository and to exchange such objects between repositories and / or the users. The METS[2] standard was designed to comply the OAIS 'Information Package' .

[2] Metadata Encoding and Transmission Standard

A Mets Document

A typical METS document consists of up to 6 sections. All sections are optional except the Structural Map section which is mandatory.

	Description	Format
Header	Dates (update, creation), status, author, role	
Descriptive Metadata	No vocabulary or syntax for encoding descriptive metadata	Any form, Dublin Core, MARC, MODS, EAD etc.
Administrative Metadata	No vocabulary or syntax for encoding administrative metadata.	
- Technical metadata	dito	Any form, e.g. MIX, FITS, PREMIS Object Metadata …
- Source metadata	dito info (descriptive, rights, technical) about the analog source used to generate the digital object	Any form, e.g. PREMIS
- Rights metadata	dito	Any form, e.g. indecs, copyrightMD, PREMIS Rights Metadata
- Digital provenance metadata	dito	Any form, e.g. PREMIS Event Metadata
File Section	All files that comprise the content of the digital entity. Files are ordered in groups (tiff, jpeg etc.) File element may refer to an external file.	
Structural Map section	Specifies the structure of the digital entity. Specifies how the files fit into this structure. More than one structure possible, e.g. logical and	

	physical.	
Behaviour Section	List all dissemination behaviours	

2.3 PREMIS

PREMIS is the acronym for **Pre**servation **M**etadata: **I**mplementation **S**trategies. The Data Dictionary of PREMIS defines preservation metadata and provides an XML schema[3]. The PREMIS Data Model[4] defines Intellectual Entities, Objects, Rights, Agents and Events.

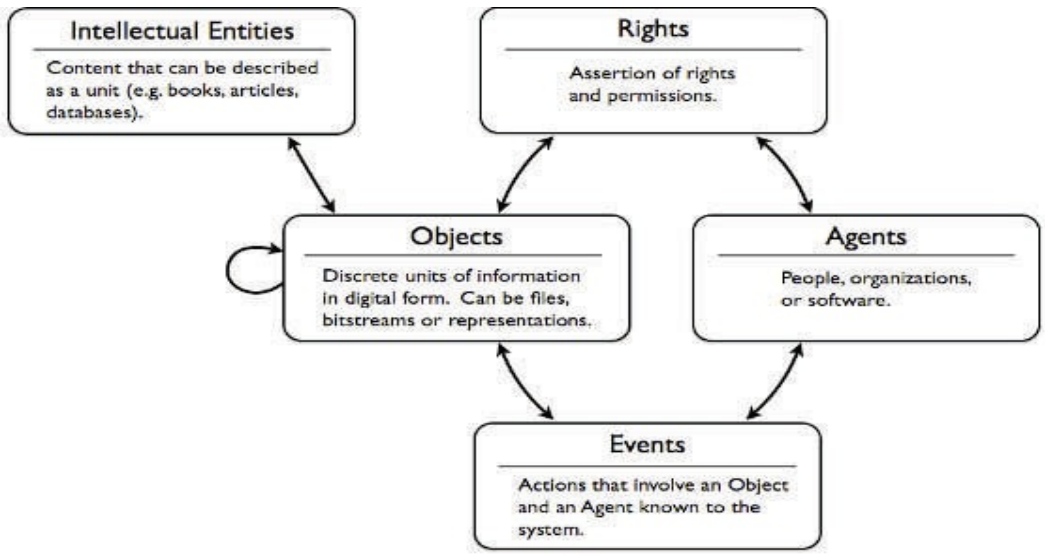


Figure 2.3-1: The Premis Data Model

An Object in PREMIS has three subtypes: file, representation and bitstream.

[3] http://www.loc.gov/standards/premis/premis.xsd
[4] http://www.loc.gov/standards/premis/

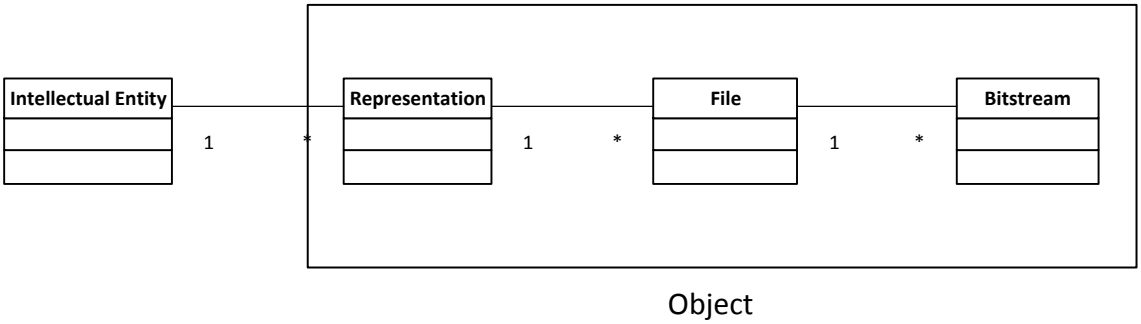
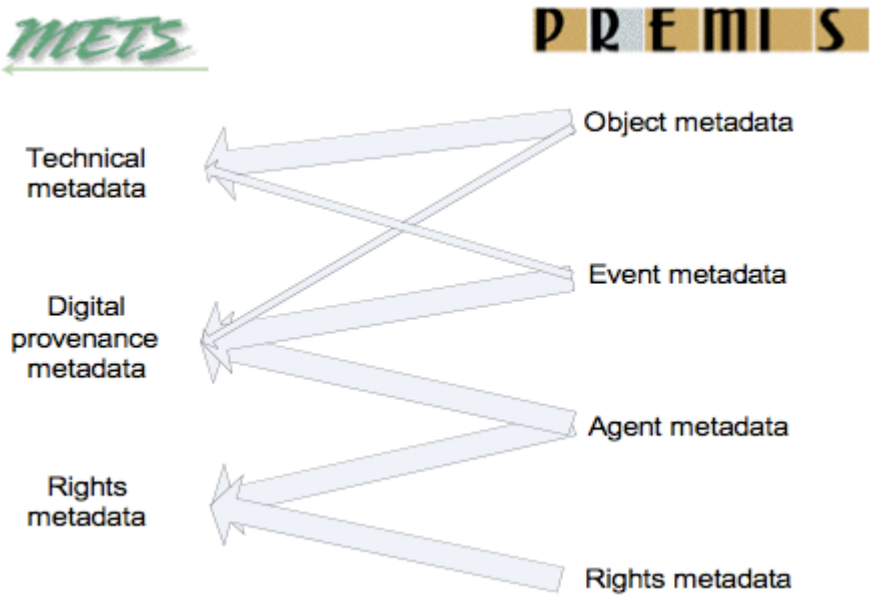


Figure 2.3-2: The PREMIS Data Model with Intellectual Entity and the Object with its subtypes: Representation, File and Bitstream.

The Intellectual Entity is e.g. a book, a map, photograph or database etc. The Representation of an IE is a set of Files including structural Metadata (e.g. described by METS). A File is a named and ordered sequence of bytes known by the operating system that can be written, read and copied. Bitstreams represent data within a file e.g. a jpeg in a PDF document, audio data within a WAVE file or graphics within a Word document.

2.4 METS & PREMIS

METS as an XML container for structuring metadata in different formats is often used in conjunction with the PREMIS standard for preservation metadata. But one has to consider the existing overlap of the METS and PREMIS definitions. There are a few documents available describing best practices and guidelines of how to use PREMIS within METS, see for example [5].



[5] http://www.loc.gov/standards/premis/guidelines-premismets.pdf

Figure 2.4-1: Mapping PREMIS entities to METS metadata sections. Thick arrows show applicable subsection in METS for the named PREMIS entities; the thin arrow shows links from one PREMIS entity to another METS subsection. (Graphic is taken from)

The following 13 checkpoints may help when dealing with METS and PREMIS:

1. How does the profile relate to other METS profiles
2. What schemas (PREMIS, MOS, MIX) are used and where are they located
3. What controlled vocabularies for PREMIS semantic units are used and where are they located?
4. Is PREMIS information wrapped into or referenced from the METS document?
5. Is PREMIS information bundled or distributed in several places in the METS document?
6. IS PREMIS information placed in separate amdSec elements or amdSec subelements?
7. Is technical metadata recorded in seperate techMD sections or with PREMIS objectCharacteristicExtension?
8. What PREMIS semantic units does the profile require or recommend?
9. Are relationships between objects expressed using METS div elements, PREMIS relationships, or both?
10. What level of object does PREMIS information describe?
11. How are PREMIS linking identifiers, IDREFs, and PREMIS identifiers used?
12. How are PREMIS-METS redundancies handled?
13. What metadata tools or applications are used?

For a detailed discussion of these checkpoints (with examples) please refer to [6]

[6] http://www.loc.gov/standards/premis/premis_mets_checklist.pdf

3 Digital Object Model of SCAPE Repositories

The SCAPE partners are using different repository implementation such as Rosetta (ExLibris), RODA (Keeps), eSciDoc (FIZ Karlsruhe) and DOMS (SB). All of these repositories do have their own Digital Object Model. We will briefly describe these models in the following section.

3.2 Rosetta

Ex Libris Rosetta is a digital-object preservation solution that conforms to the ISO-recognized Open Archival Information System (OAIS).

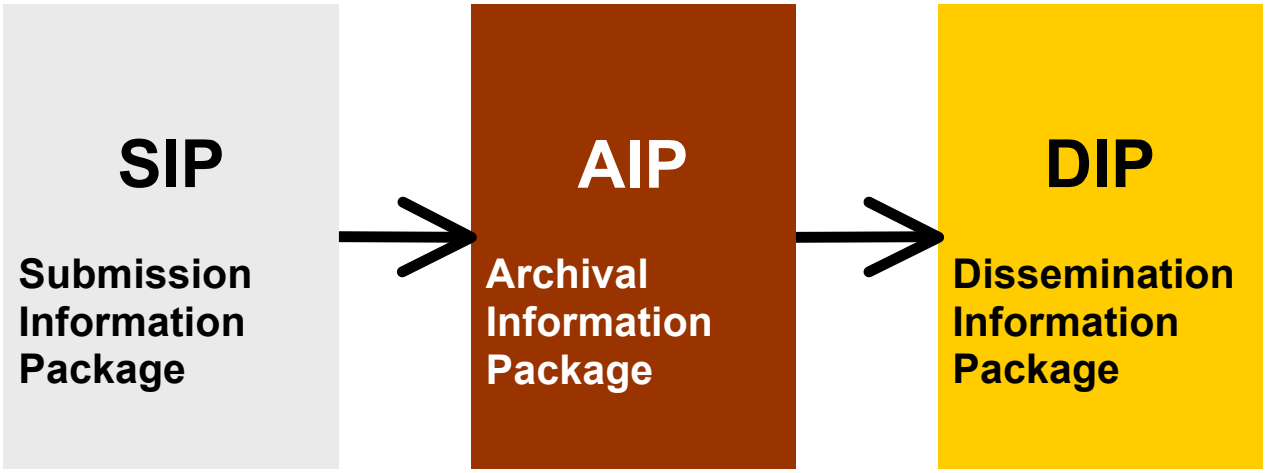
Rosetta allows the SIP and the DIP to have a variety of formats and structures and provides an SDK to support this.

The following chapter describes the AIP, which is stored in a METS XML file in Rosetta's Permanent Repository module. Each AIP describes one IE (intellectual entity).

The METS XML is generated in the Staging module during the SIP processing. During processing, the IE information is kept and managed in the database. By the time the SIP is moved to the permanent repository, the METS XML contains all the information regarding the IE, collected from the different database tables.

The information on the METS XML can be reloaded back into the database when the IE is brought from the permanent repository for maintenance (preservation actions, adding representations, and so forth).

The following diagram shows the flow between the three types of information packages:

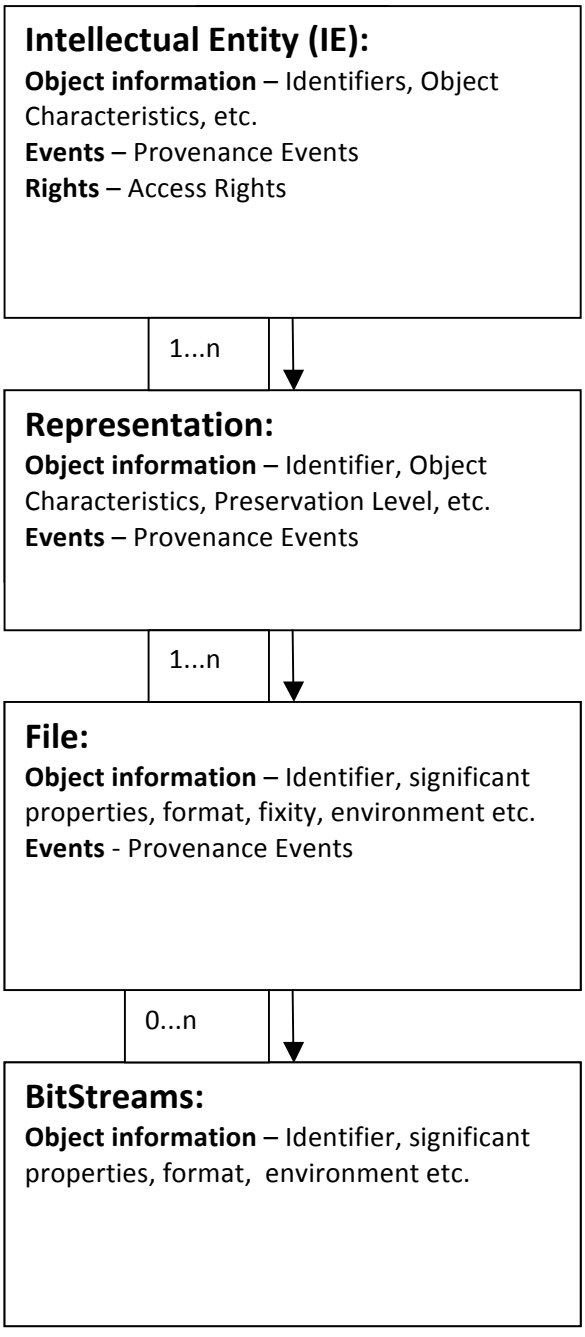


The AIP is stored in a METS XML file and can be viewed in the Permanent Repository module of Rosetta. Each AIP relates to one IE (intellectual entity).

3.2.1 Structure

Rosetta AIP data model that is based on the PREMIS reference model covers four levels of objects. Further information related to the PREMIS reference model can be found at: http://www.loc.gov/standards/premis/

The following figure illustrates the four entities of the AIP data model:



The entities in the data model are defined as follows:

- **Intellectual Entity** – A set of content that is considered a single intellectual unit for purposes of management and description – for example, a particular book, map, photograph, or database. An intellectual entity may have one or more digital representations.
- **Representations** – A Representation is the set of files, including structural metadata, needed for a complete and reasonable rendition of an intellectual entity. There can be more than one representation for the same intellectual entity. For example, a journal article may be complete in one PDF file and this single file will then constitute the representation. However, another journal

article may consist of one SGML file and two image files. In this case, these three files will constitute the representation. A third article may be represented by one TIFF image for each of 12 pages plus an XML file of structural metadata showing the order of the pages. In this case, 13 files will constitute the representation. (PREMIS data dictionary, p. 14)

- **Files** – A file is a named and ordered sequence of bytes that is known by an operating system. A file can be zero or more bytes and has a file format, access permissions, and file system characteristics such as size and last modification date.
- **Bitstreams** – A bitstream is contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes. A bitstream cannot be transformed into a standalone file without the addition of file structure (headers, and so forth) and/or reformatting to comply with a particular file format.

  A bitstream is defined in the PREMIS data model as a set of bits embedded within a file. This differs from common usage, where a bitstream could, in theory, span more than one file.

  A good example of a file with embedded bitstreams is a TIFF file containing two images.

  According to the TIFF file format specification, a TIFF file must contain a header that includes information about the file. It may then contain one or more images. In the data model, each of these images is a bitstream and can have properties such as identifiers, location, inhibitors, and detailed technical metadata (for example, color space).

  Some bitstreams have the same properties as files and some do not. The image embedded within the TIFF file clearly has properties that are different from the file itself. However, three TIFF files can also be aggregated within a larger TAR file. In this case, the three TIFF files are filestreams, but they have all the properties of TIFF files. [7]

  Rosetta bitstream functionality is limited to filestream only. Real bitstreams (embedded objects within a file) are functionally not supported however from a Data Model perspective, the Data Model serves both types of bitstreams

Rosetta uses METS as a container for the IE as an AIP. Further information related to the METS reference model can be found at [8]

The METS schema contains three types of metadata: Descriptive, Administrative, and Structural Map.

The following table illustrates which metadata type is applicable to each object type:

[7] http://www.loc.gov/standards/premis/v2/premis-2-1.pdf
[8] http://www.loc.gov/standards/mets/

	Descriptive	Administrative	Structural Map
IE	√	√	
Representation		√	√
File		√	
Bitstream		√	

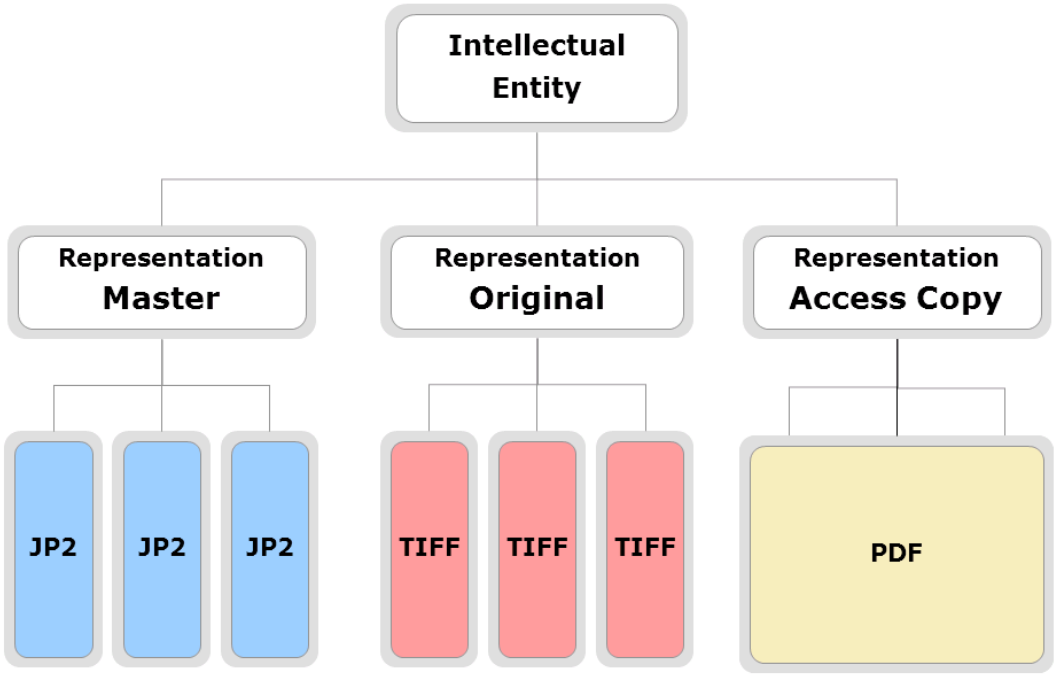


Figure 3-1: An Illustration of a Mets Container in Rosetta

3.3 RODA

RODA is an open source digital repository with long-term preservation and authenticity as its primary objectives. Created by the Portuguese National Archives in partnership with the University of Minho, it was designed to support the most recent archival standards and become a trustworthy digital repository.

RODA has been developed to be a complete digital repository providing functionality for all the main units that compose the OAIS reference model. As an OAIS compliant repository, RODA defines a Submission Information Package (SIP), an Archival Information Package (AIP) and a Dissemination Information Package (DIP) format for ingest, archival and dissemination of information, respectively.

In the following sections RODA SIP and AIP are further described.

3.3.1 SIP

New data is submitted to the repository in the shape of Submission Information Packages (SIP). When the ingest process terminates, SIPs are transformed into Archival Information Packages (AIP), i.e. the actual packages that will be kept in the repository. Associated with the AIP is the structural, technical and preservation metadata, as they are essential for carrying out preservation activities.

The SIP is composed of one or more digital representations and all of the associated metadata, packaged inside a METS envelope. Producers take advantage of a small application called RODA-in that allows them to create these packages. The structure of a SIP supported by RODA is depicted in Figure 3-2. The RODA SIP is basically a compressed ZIP file containing a METS envelope, the set of files that compose the representations and a series of metadata records. Within the SIP there should be at least one descriptive metadata record in EAD-Component[9] format.

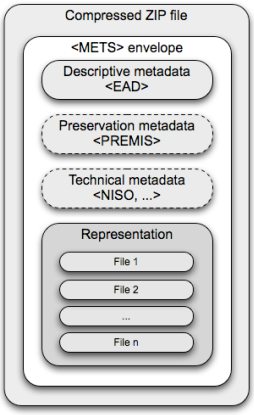


Figure 3-2: Structure of a Submission Information Package in RODA using a METS envelope

One may also find preservation and technical metadata inside a submission package, although this last set of metadata is not mandatory as is seldom created by producers. Nevertheless, it was felt important that RODA should support those additional SIP elements for special situations such as repository succession, e.g. when ingested items belong to a repository that is to be deactivated.

[9] An EAD record does not describe a single representation. In fact, EAD is used to describe an entire collection of representations. In the SIP is included only a segment of EAD that is sufficient to describe one representation, i.e. a <c> element and all its sub-elements. The team has called this subset of the EAD an EAD-Component.

The main file inside a SIP is the METS envelope file (METS.xml) that structures the SIP information. This envelope file references the descriptive metadata which can reference one or more representations that in turn reference the files. Representations can also reference preservation and technical metadata.

The following snippet is an example of a METS envelope file representing a SIP in RODA:

```
<?xml version="1.0" encoding="UTF-8"?>
<mets PROFILE="RODA_SIP" xmlns="http://www.loc.gov/METS/"
xmlns:xlin="http://www.w3.org/1999/xlink">
  <metsHdr>
    <agent ROLE="CREATOR">
      <name>RODA Common SIP Utility</name>
    </agent>
  </metsHdr>
  <dmdSec ID="EADC-WALLPAPER--1664823803">
    <mdRef LOCTYPE="URL" MDTYPE="OTHER" xlin:href="eadc5700174555992435226.tmp"
       LABEL="roda:d" CHECKSUMTYPE="MD5"
       CHECKSUM="12069DF39F92FF1682FACE1E1FC2B712"/>
  </dmdSec>
  <dmdSec ID="EADC-NASA-1691196776" GROUPID="EADC-WALLPAPER--1664823803">
    <mdRef LOCTYPE="URL" MDTYPE="OTHER" xlin:href="eadc2713347945811457232.tmp"
       LABEL="roda:d" CHECKSUMTYPE="MD5"
       CHECKSUM="EEC03168F8C3D30D03A9089F1DF31685"/>
  </dmdSec>
  <fileSec>
    <fileGrp>
      <file ID="R2012-03-28T14.44.44.49Z-F0" MIMETYPE="application/octet-stream"
            CHECKSUMTYPE="MD5" CHECKSUM="5F4CDB67BC8EA454A70FA448BDF71B31">
        <FLocat LOCTYPE="URL" xlin:href="R2012-03-28T14.44.44.49Z/F0"
            xlin:title="METS.xml"/>
      </file>
      <file ID="R2012-03-28T14.44.44.49Z-F1" MIMETYPE="image/png"
            CHECKSUMTYPE="MD5" CHECKSUM="EEF92204A161CC8C157B4D3F8B7FBA74">
        <FLocat LOCTYPE="URL" xlin:href="R2012-03-28T14.44.44.49Z/F1"
          xlin:title="magField_3d_02Left_lrg.png"/>
      </file>
      <file ID="R2012-03-28T14.44.44.49Z-F3" MIMETYPE="image/jpeg"
            CHECKSUMTYPE="MD5" CHECKSUM="829AD7D9A296FFFE86B4E01E8E6D9178">
        <FLocat LOCTYPE="URL" xlin:href="R2012-03-28T14.44.44.49Z/F3"
          xlin:title="keithburns-1920.jpg"/>
      </file>
    </fileGrp>
  </fileSec>
  <structMap>
    <div ID="Representations">
      <div ID="R2012-03-28T14.44.44.49Z"
            TYPE="roda:r:digitalized_work:image/mets+misc"
            DMDID="EADC-NASA-1691196776" xlin:label="original">
        <fptr FILEID="R2012-03-28T14.44.44.49Z-F0"/>
        <fptr FILEID="R2012-03-28T14.44.44.49Z-F1"/>
        <fptr FILEID="R2012-03-28T14.44.44.49Z-F3"/>
      </div>
    </div>
  </structMap>
</mets>
```

After a successful validation, the SIP is then taken apart and each of its constituents is integrated into the repository. After this procedure, the SIP becomes an AIP and is ready to be disseminated to potential consumers that have clearance to access that information.

3.3.2 AIP

Unlike the SIP, RODA's AIP does not rely on METS to structure the information, it rather uses Fedora Commons RDF linking mechanism to keep the reference integrity between the several AIP constituents.

RODA's content model is atomistic and very much PREMIS-oriented (Figure 3-3). Each intellectual entity is described by an EAD-component metadata record (DO nodes in Figure 3-3). These records are organized hierarchically in order to constitute a full archival description but are kept separately within the Fedora Commons content model.

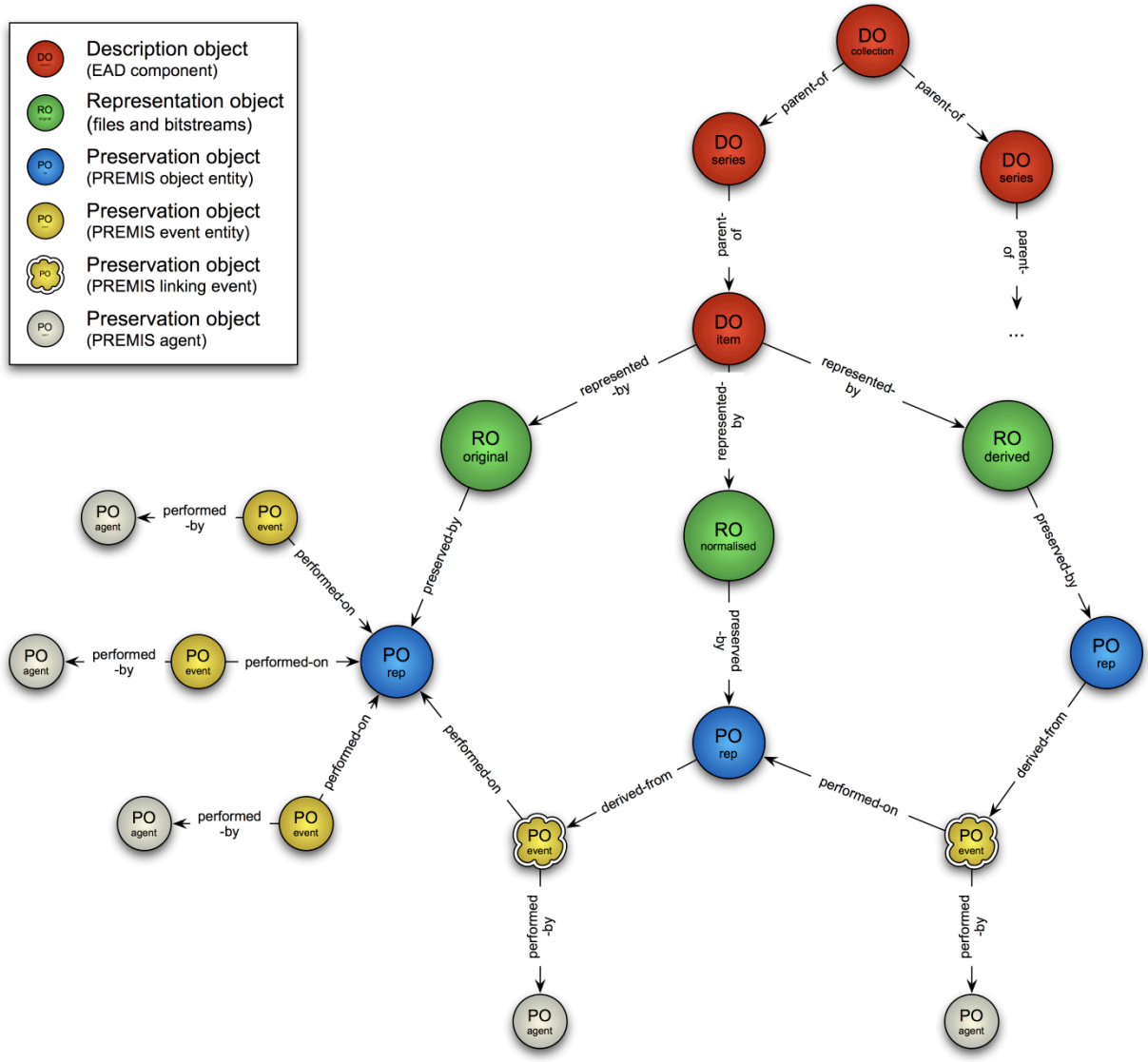


Figure 3-3: RODA's molecular content model featuring EAD-components, digital

Additionally, each leaf record (i.e. a file or an item) is linked to a representation object (RO nodes in the figure), i.e. a fedora object that embeds all the files and bit-streams that compose the digital

representation. Finally, each of these objects are linked together by a set of PREMIS entities that maintain information about the digital object's provenance and history of events (PO nodes).
Each preservation event that takes place inside the repository is recorded as a new preservation-event node (i.e. PO event nodes in the figure). Special events, like format migrations, establish relationships between two preservation-representation nodes. These are called linking events. Each preservation event is executed by an agent, whether this is a system user or an automatically triggered software application. The agent that triggered the event is recorded in PO agent nodes.

3.4 eSciDoc

eSciDoc is a joint project of the Max Planck Society and FIZ Karlsruhe. It primarily focuses on the support of the scientific life cycle but can be used for other approaches as well. It comprises core functionality including a Fedora Commons repository, a set of complementing services, and application build on top of the infrastructure and the services that enable innovative eScience scenarios. Scientists, librarians, and software developers can work with research data, create novel forms of publications, and establish new ways of scientific and scholarly communication.[10]
eSciDoc is able to use either managed content or external referenced content.

Even though eSciDoc covers many disciplines and use cases in the field of eScience, eSciDoc has only three generic object patterns:

- Context (administrative container)
- Container (aggregations, e.g., collections, bundles)
- Item

Each content resource (Item or Container) is maintained in a single administrative Context. The two content resources, Item and Container are defined as follows:

- An *Item* resource consists of metadata records (e.g. eSciDoc publication metadata, MODS record, Dublin Core record) and optionally of *Components* that represent the actual content (e.g. PDF file, JPEG file, XML file).
- A *Container* resource is an aggregation of other resources that allows for aggregating other items or containers. Like the Item resource, Container can be described by multiple metadata records.

Items and Containers are very generic resources and they do not speak for themselves about the content they represent or about their own structure e.g. what kind of metadata may be associated with them, what kind of members a container aggregates, or what kind of resources they represent semantically. Therefore, eSciDoc logical data model introduced the concept of a Content model. Each Item or a Container has to claim that it is an instance of exactly one content model. There is no limitation on the number of instances that may claim to be of a specific content model.

Content model defines in general:

- the type and structure of the content resources (Item, Container, Components)
- a set of services that may be associated with the content resources

[10] https://www.escidoc.org/

The following graphic illustrates the eSciDoc object patterns and their relationships for a specific example with a Context “My Science Lab”:

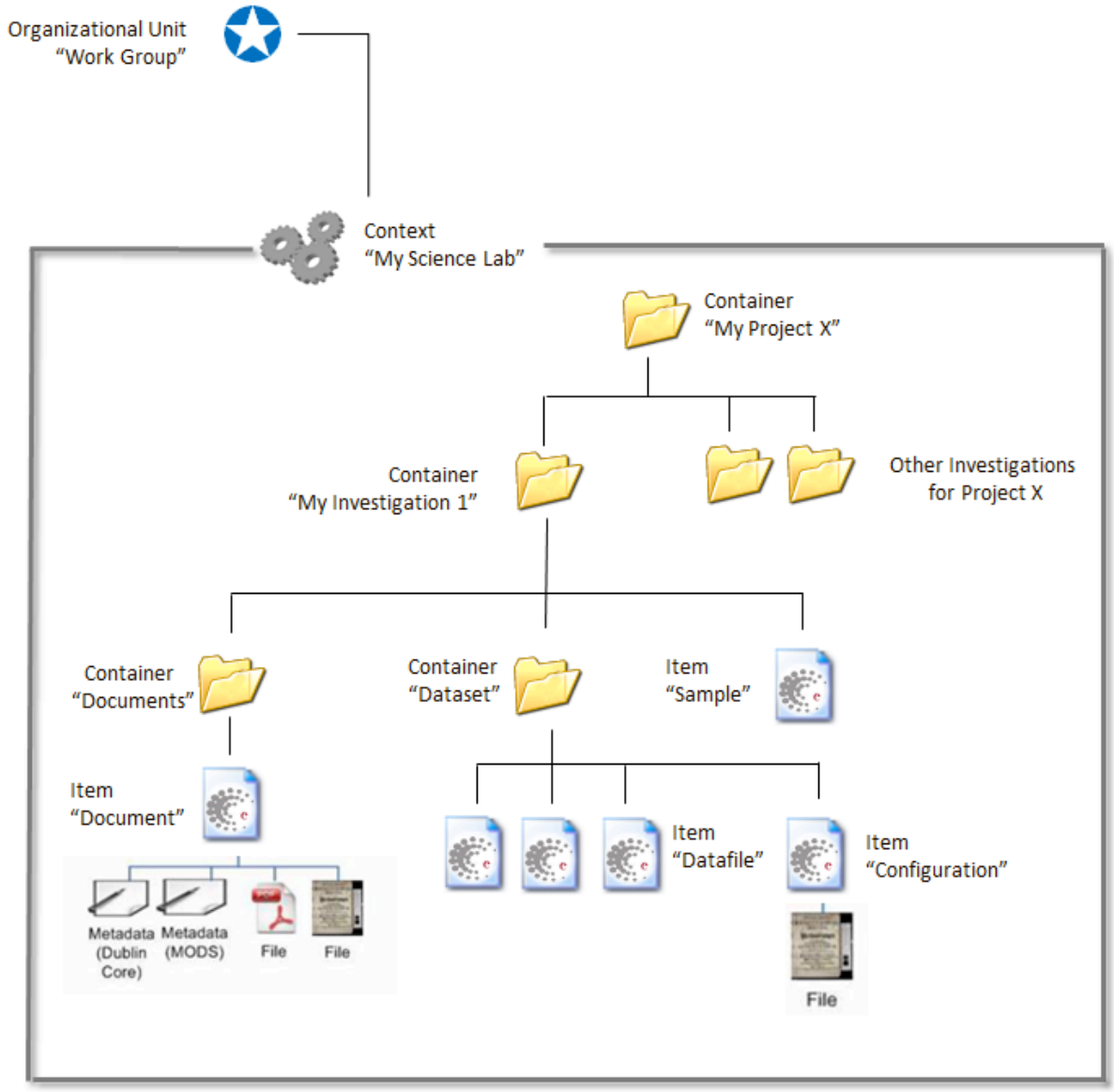


Figure 3-4: sample view on the generic object patterns and their relationship to each other of eSciDoc for a Context “My Science Lab”.

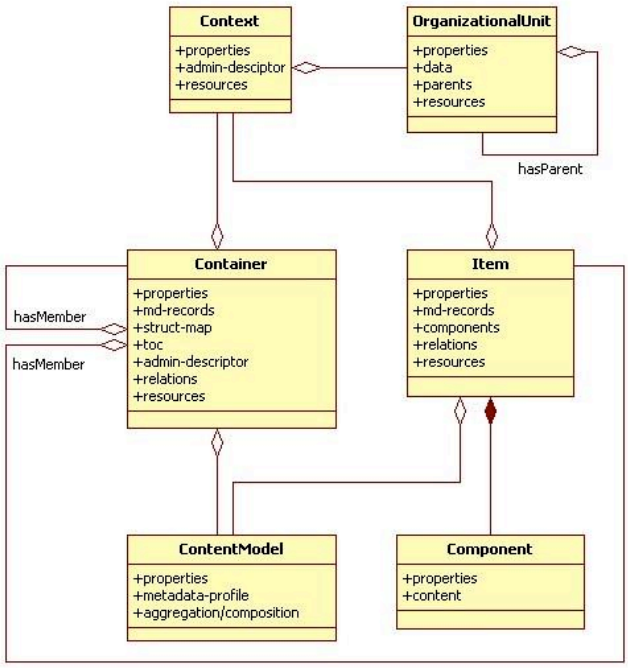


Figure 3-5: A technical view of the eSciDoc Data Model

eSciDoc needs to map the METS profile to the existing eSciDoc data model. One eSciDoc item may represent an Intellectual Entity and the eSciDoc components may be mapped to the representations. It is also possible to have an eSciDoc Container represent an Intellectual Entity and have the eSciDoc Items map to the representations and have its eSciDoc Components map to files.

A possible mapping of the eSciDoc Data Model to PREMIS terms:

eSciDoc	PREMIS
Container	Intellectual Entity
Item	Representation / Intellectual Entity
Component	Representation / File / Bitstream

4 The SCAPE Digital Object Model

A possible Digital Object model is outlined in the following section as discussed during the Den Haag Meeting 27-29th of February with members of FIZ Karlsruhe and ExLibris.
In this model each Intellectual Entity will be represented by one METS file, and each Representation File and Bitstream will be described by technical metadata.

4.1 Requirements METS

In SCAPE we define one METS profile for all use cases and OAIS Information Packages.
The different optional and mandatory properties of the METS profile for SIP and AIP will be indicated in the profile itself. For SIP a minimal set of elements are mandatory and for the AIP we will have all elements that are needed for preservation are mandatory.

Formats for all relevant Metadata in the SCAPE METS profile

- Descriptive Metadata (Dublin Core, EAD, MODS)
- Technical Metadata (PREMIS, MIX, VideoMD, AudioMD, textMD, JHOVE)
- Digital Provenance Metadata (PREMIS event and agents)
- Source Metadata (contains descriptive, rights or technical metadata)
- Rights Metadata (PREMIS rights)

In order to map PREMIS entities to METS we follow the guidelines described in "Guidelines for using PREMIS with METS for exchange" [11]. The following table illustrates the mapping:

PREMIS	METS
premis:object	techMD or digiProvMD
Premis:event	digiProvMD
premis:rights	rightsMD
premis:agent	digiProvMD or rightsMD

A generic METS profile that uses PREMIS is available at [12].
The following METS XML snippet shows the basic structure of a METS XML document for one Intellectual Entity "fiz.karlsruhe.09601" with two objects. Note: the sections are empty.

```
<mets:mets PROFILE="SCAPE" OBJID="fiz.karlsruhe.09601"
xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/mets.xsd">
<mets:dmdSec ID="DC"></mets:dmdSec>
<mets:amdSec>
       <mets:techMD ID="object1"></mets:techMD>
       <mets:rightsMD ID="rights1"></mets:rightsMD>
       <mets:sourceMD ID="source1"></mets:sourceMD>
       <mets:digiprovMD ID="event1"></mets:digiprovMD>
       <mets:techMD ID="object2"></mets:techMD>
       <mets:rightsMD ID="rights2"></mets:rightsMD>
       <mets:sourceMD ID="source2"></mets:sourceMD>
       <mets:digiprovMD ID="event2"></mets:digiprovMD>
</mets:amdSec>
<mets:fileSec></mets:fileSec>
<mets:structMap></mets:structMap>
</mets:mets>
```

The following examples shows the usage of PREMIS within the METS techMD section:

[11] http://www.loc.gov/standards/premis/guidelines-premismets.pdf
[12] http://www.loc.gov/standards/premis/louis-2-0.xml

```
<mets:techMD ID="object1">
<mets:mdWrap MDTYPE="PREMIS:OBJECT">
<mets:xmlData>
<premis:object xsi:type="premis:file"
xsi:schemaLocation="info:lc/xmlns/premis-v2
http://www.loc.gov/standards/premis/v2/premis-v2-0.xsd">
       <premis:objectIdentifier></premis:objectIdentifier>
       <premis:preservationLevel></premis:preservationLevel>
       <premis:significantProperties></premis:significantProperties>
       <premis:objectCharacteristics></premis:objectCharacteristics>
       <premis:originalName></premis:originalName>
       <premis:storage></premis:storage>
       <premis:environment></premis:environment>
       <premis:relationship></premis:relationship>
       <premis:linkingEventIdentifier></premis:linkingEventIdentifier>
       <premis:linkingIntellectualEntityIdentifier></premis:linkingInte
llectualEntityIdentifier>
</premis:object>
</mets:xmlData>
</mets:mdWrap>
</mets:techMD>
```

4.2 Example METS Profile

In order to guarantee easy interchangeability of Platform implementations for SCAPE a METS profile with the following characteristics can be used:

A METS file adhering to this profile

- may contain a “<metsHdr>” element
- must contain one and only one “<dmdSec>” element
- must use the “<mdWrap>” or <mdRef> element for metadata
- must have one “<amdSec>” element [not for SIP]
- must have one <techMD> for each file [not for SIP]
- may contain <rightsMD>, <sourceMD>, <digiprovMD> for each file.
- must have a “<fileGrp>” element for each Representation
- must have a “<file>” element containing an “<FLocat>” element with location information as a URL in the “xlink:href” attribute
- must have a “<structMap>” element.
- must not have a “<structLink>” element
- must not have a “<behaviourSec>” element
- must wrap descriptive, technical, rights, source and provenance metadata in a “<mdWrap>” or <mdRef> element composed of PREMIS elements.
- must only use descriptive metadata composed of Dublin Core terms
- must only use technical metadata composed of NISO Z39.87 terms for digital still images (MIX), TextMD for textual representations [...]
- must only use provenance metadata composed of PREMIS events

- must only use source metadata composed of Dublin Core terms
- must only use rights metadata composed of PREMIS rights

The following figure shows the SCAPE METS profile with optional (grey), mandatory (blue) and reference implementation (green) elements.

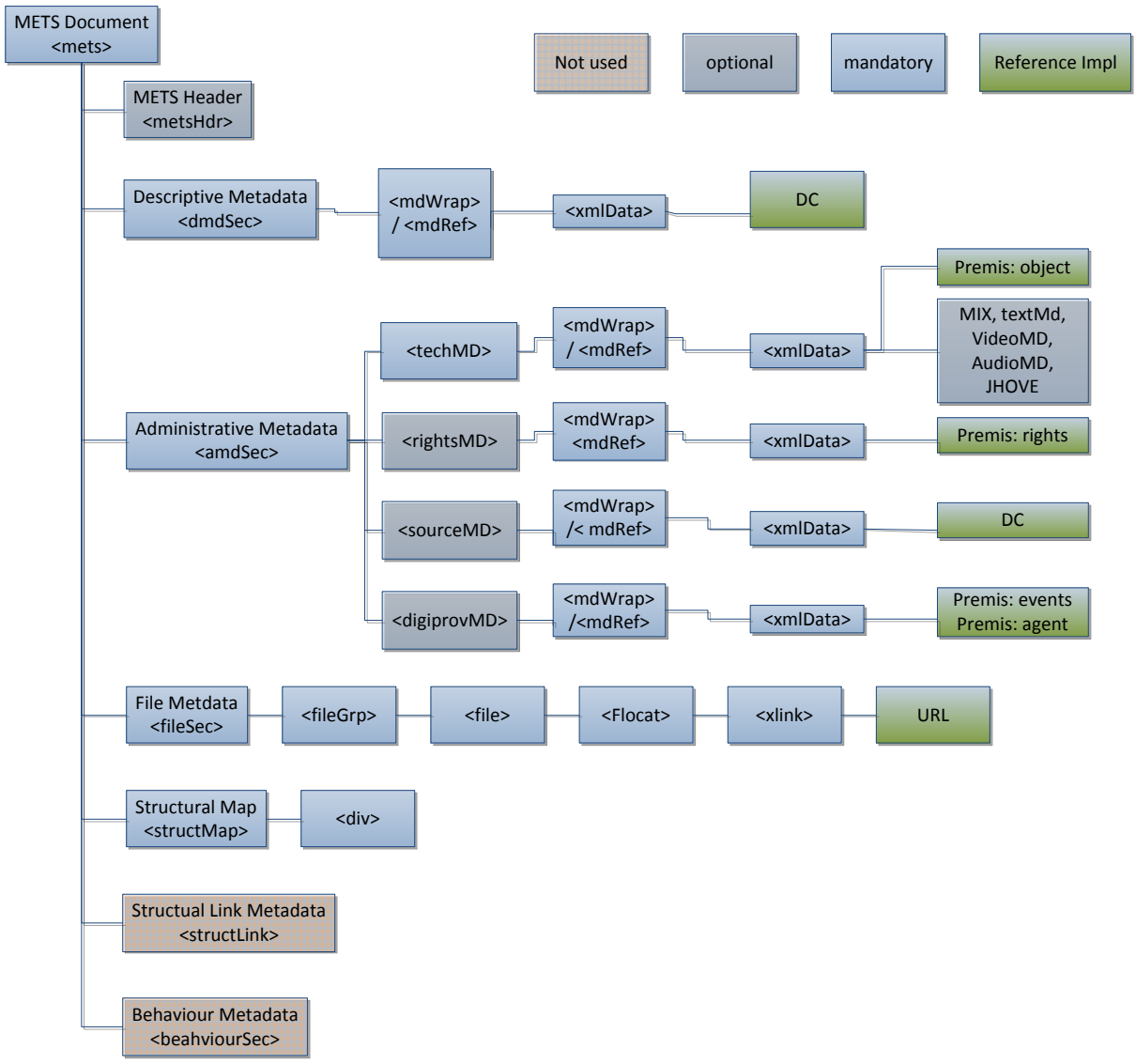


Figure 4-1: SCAPE METS Profile

4.3 METS Structmap

The METS document incudes exactly one <structmap> element which is used to map the structure of the intellectual entity into a flat METS XML representation. The map consists of nested elements that correspond to the structure of a Intellectual Entity , i.e. Intellectual Entity, Representation, File, Bitstream. So there is exactly one element on the top level corresponding to the Intellectual entity. This div contains nested elements for each representation, that are linked via identifiers to the metadata describing the representation. Each representation contains a set of nested elements for each File, linking to it's technical metadata. Each of the elements on the file level contain exactly one <fptr> element for the identification of the File. The -element on the file level in turn can contain nested elements for each Bitstream in the File.

4.3.1 StructMap example

```
<div id="1" type="IntellectualEntity" label="entity_1" dmdid="dmd-1">
    <div id="2" type="Representation" label="rep-1">
        <div id="3" type="techMD" admid="tech-1"/>
        <div id="4" type="rightsMD" admid="reghts-1"/>
        <div id="5" type="digiProvMD" admid="digiprov-1"/>
        <div id="6" type="sourceMD" admid="source-1" order="0"/>
        <div id="7" type="File">
            <fptr fileid="file-1"/>
            <div id="8" type="techMD" admid="tech-2"/>
            <div id="9" type="Bitstream" >
                    <div id="10" type="techMD" admid="tech-3"/>
            </div>
        </div>
     </div>
</div>
```

4.4 METS and PREMIS identifiers

Each METS document must be assigned a persistent and unique identifier. This identifier must be locally and globally resolvable. Possible identifier schemes are: OCLC Purls[13], CNRI Handles[14], DOI[15] etc. But it is also possible to use just UUID[16] to assign a unique - but not global - identifier. A further requirement is that the identifier must be a string. Further Information can be found at [17] .

For a SIP the METS identifier is optional assuming that the repository will assign an identifier on submission. The identifier will be recorded in the OBJID attribute of each METS document.

The PREMIS identifiers such as objectIdentifier, agentIdentifier, permissionStatementIdentifier must be consistent within each METS document, but don't have to be globally or locally resolvable.

[13] http://purl.org/docs/index.html
[14] http://www.handle.net/
[15] http://www.doi.org/
[16] http://www.itu.int/ITU-T/studygroups/com17/oid.html
[17] http://en.wikipedia.org/wiki/Universally_unique_identifier

4.5 Extension Schemas

- DC – Dublin core
  - May be included in the dmdMD element. Each METS document must contain a dmdSec describing the entire package.
  - http://dublincore.org/schemas/xmls/

- PREMIS
  - See table in section 8.1 for the usage of PREMIS in METS.
  - http://www.loc.gov/standards/premis/v2/premis-v2-1.xsd

- Technical Metadata for Digital Still Images (MIX) - NISO Data Dictionary
  - May be included in the techMD element
  - http://www.loc.gov/standars/mix/mix.xsd

- textMD – Text Metadata Schema
  - May be included in the techMD element
  - http://dlib.nyu.edu/METS/textmd.xsd

- VideoMD – Video Technical Metadata Extension Schema
  - May be included in the techMD element
  - http://lcweb2.loc.gov/mets/Schemas/VMD.xsd

- AudioMD – Audo Technical Metadata Extension Schema
  - May be included in the techMD element
  - http://lcweb2.loc.gov/mets/Schemas/AMD.xsd

- JHOVE XML handler output schema
  - May be included in the techMD element
  - http://hul.harvard.edu/ois/xml/xsd/jhove/jhove.xsd
- FITS File Information Tool Set
  - May be used in the techMD element
  - http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd

4.6 Requirements for the OAIS Information Packages

We need to define the three Information Packages described by OAIS for SCAPE

4.6.1 Definition of a SIP

The most simplest SIP contains a header and content (files). A realistic example for a SIP is illustrated in the RODA section of this document. This definition of a SIP can be used for the SCAPE Loader Application as well. The following figure shows the minimum definition of a SIP METS profile used within SCAPE:

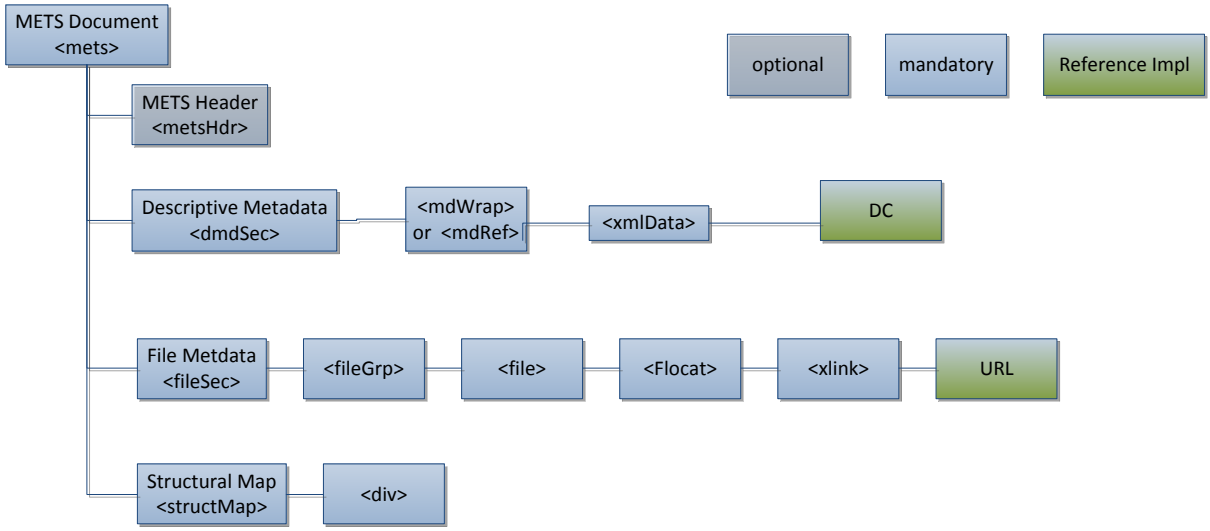


Figure 4-2: SCAPE METS profile of a SIP

4.6.2 Definition of a AIP

An AIP is an archived Intellectual Entity and should therefore contain technical metadata and digital preservation metadata. An AIP may also contain information about the preservation plan associated with the Intellectual Entity. Please refer to section 4.2 for further information about the METS profile of an AIP.

4.6.3 Definition of a DIP

A Dissemination Information Package (DIP) in SCAPE will contain the same information as an Archival Information Package (AIP).

4.7 Preservation Plans

Preservation Plans will not be described by this SCAPE METS profile but do have their own XML schema definition[18], although it can be wrapped in a METS container. If a Plan gets executed the provenance information and the information about the plan executed gets stored in the digital provenance section in the AIP of the digital object. The Plan itself may be stored as an AIP in the repository.

4.8 Summary

The following graphic illustrates the OAIS Information packages and the relationship with an Intellctual Entity:

[18] http://www.ifs.tuwien.ac.at/dp/plato/schemas/plato-3.0.1.xsd - this definition will change in the future.

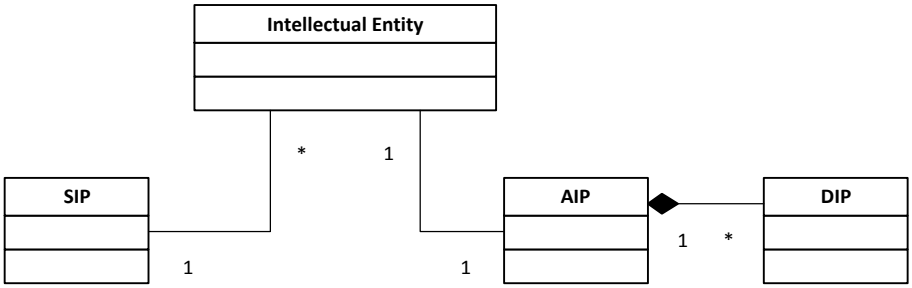


Figure 4-3 The possible Relation of an Intellectual Entity with a SIP, AIP and DIP

One SIP can hold several Intellectual Entities. One Intellectual Entity is represented by one AIP. A DIP is either a one-to-one representation of an AIP or can be just part of a AIP.

The following table summarizes the definition of a METS profile of a SIP and AIP:

METS section	AIP	SIP
<metsHdr>	optional	optional
<dmdSec>	mandatory	mandatory
<amdSec>	mandatory	-
<fileSec>	mandatory	mandatory
<structMap>	mandatory	mandatory
<structLink>	-	-
<behaviourSec>	-	-

5 Conclusion

We described a Digital Object Model of SCAPE based on METS and PREMIS. We explained a METS profile that must be used by all SCAPE partners in order to prevent a situation where a partner will realize only in the last year that he cannot integrate with SCAPE because his data model is something totally different. Beside that we defined the OAIS Information packages used by the repositories and other systems or users that interact with the repository (e.g. SCAPE Watch component).

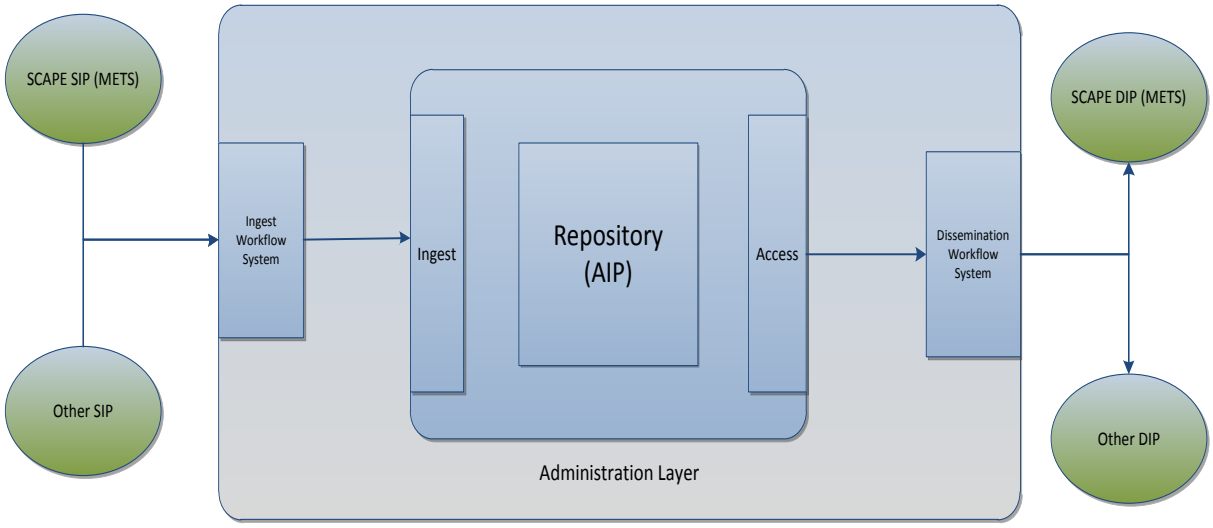


Figure 5-1: Repository System showing SCAPE SIP and DIP

6 Glossary

Representation (PREMIS term)

The set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity. For example, a journal article may be complete in one PDF file; this single file constitutes the representation. Another journal article may consist of one SGML file and two image files; these three files constitute the representation. A third article may be represented by one TIFF image for each of 12 pages plus an XML file of structural metadata showing the order of the pages; these 13 files constitute the representation. From Introduction and Supporting Materials from PREMIS Data Dictionary, p. 7.[19]

Metadata

Information about an analog or digital object, a component of an object, or a coherent collection of objects. Metadata describing digital content is often structured (e.g., with tagging or markup) and it may be embedded within a single file, incorporated within the "packaging" that is associated with a group of files (e.g., METS), placed in a related external file (e.g., XMP sidecar file), or in a system external to the digital file (e.g., a database) to which the digital file or files are linked via a unique key or association.[20]

Metadata, administrative

Metadata used for the management of digital content, such as information about rights and permissions (see Metadata, rights) as well as other facts about a given digital object. Some speakers define *administrative metadata* to include technical metadata (see Metadata, technical), source metadata (see Metadata, source), and process metadata (see Metadata, process).[21]

Intellectual entity (PREMIS term)

A set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities; for example, a Web site can include a Web page; a Web page can include an image. An Intellectual Entity may have one or more digital representations. From Introduction and Supporting Materials from PREMIS Data Dictionary, p. 6.[22]

File (PREMIS term)

file is a named and ordered sequence of bytes that is known by an operating system. A file can be zero or more bytes and has a file format, access permissions, and file system characteristics such as size and last modification date . . . . Files can be read, written, and copied. Files have names and formats." *Introduction and Supporting Materials from PREMIS Data Dictionary* (p. 7)[23]

[19] http://www.digitizationguidelines.gov/term.php?term=representation
[20] http://www.digitizationguidelines.gov/term.php?term=metadata
[21] http://www.digitizationguidelines.gov/term.php?term=metadataadministrative
[22] http://www.digitizationguidelines.gov/term.php?term=intellectualentity
[23] http://www.digitizationguidelines.gov/term.php?term=digitalfile

Bitstream

Contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes. Generally speaking, a bitstream cannot be transformed into a standalone file without the addition of file structure (headers, etc.) and/or reformatting the bitstream to comply with some particular file format. This definition is derived from the data model outlined in *Introduction and Supporting Materials from PREMIS Data Dictionary*, p. 7, illustrated by the example of a TIFF file that contains embedded bitstreams representing raster images together with header that presents some information about the file. The authors of the PREMIS definition note that their definition is limited to sets of bits embedded within a file and they call attention to an alternate usage that defines *bitstream* as an entity that could span more than one file.[24]

Archival Information Package (AIP)

An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.[25]

Dissemination Information Package (DIP)

The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS.[26]

Submission Information Package (SIP)

An Information Package that is delivered by the Producer to the OAIS for use in the construction of one or more AIPs.[27]

7 List of Figures

[24] http://www.digitizationguidelines.gov/term.php?term=bitstream
[25] Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1 BLUE BOOK
[26] Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1 BLUE BOOK
[27] Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1 BLUE BOOK