Australian Partnership for Sustainable Repositories

PREMIS Requirement Statement

Project Report

Bronwyn Lee, Gerard Clifton and Somaya Langley

National Library of Australia
July 2006

Table of contents

Report and recommendations

This report contains the results of the PRESTA - PREMIS Requirements Statement project undertaken by the National Library of Australia from December 2005 to June 2006 for the Australian Partnership for Sustainable Repositories (APSR).

1. Australian Partnership for Sustainable Repositories (APSR)

APSR aims to establish a centre of excellence for the management of scholarly assets in digital format.

It has an overall focus on the critical issues of the access continuity and the sustainability of digital collections. It is building on a base of demonstrators in developmental repositories within partner institutions. It is contributing to national strength in this area by encouraging the development of skills and expertise and providing coordination throughout the sector. It is actively providing international linkages and national services.

APSR is supported by the Systemic Infrastructure Initiative as part of the Australian Government's Backing Australia's Ability - An Innovative Action Plan for the Future. The current partners are Australian National University, National Library of Australia, University of Queensland, University of Sydney, University of Melbourne, University of Technology Sydney and the Australian Partnership for Advanced Computing.

2. PREMIS Requirement Statement project (PRESTA)

"PRESTA - PREMIS Requirement Statement" is one of the projects of APSR. It has as its aims:

> to specify requirements for the collection of metadata needed for preservation management purposes and help these to be applied to selected repository implementations of APSR partners.

This report covers work done at the National Library of Australia during the 6 month period funded by the APSR from December 2005 to June 2006. The National Library of Australia has two digital repositories: PANDORA, Australia's web archive, and a digital repository for storing its own digital collections which is managed using a system developed inhouse called the Digital Collections Manager (DCM). The National Library of Australia has been actively involved in national and international digital preservation initiatives and information about its activities can be found on the Digital Preservation part of the National Library of Australia's website.

The "selected repository implementations" studied were the Australian National University's (ANU) Demetrius repository based on DSpace and the University of Queensland's (UQ) eScholarship repository based on Fez and Fedora.

3. Scope of the project

The original draft workplan implied an expectation that functional specifications for collection of metadata during submission and ingest would be written which the repositories would then implement. However it was felt not to be appropriate for the selected repositories, because their systems were already implemented with established business and submission models. It was decided there would be more emphasis on *what* metadata was collected than *how* it was collected. The project would specify the metadata needed for preservation purposes, identify metadata that was not currently being collected and make recommendations for enhancements, but leave decisions on how to implement those enhancements to the repositories themselves.

Use cases *were* written for preservation events and their metadata, as this was an area lacking in both ANU and UQ repositories which has not been covered elsewhere. The other significant gap, preservation risk monitoring, is being addressed in the AONS (Automatic Obsolescence Notification System) project.

Although the title of the project was "PREMIS Requirement Statement" the project did not confine itself to PREMIS but considered *all* metadata, including PREMIS, necessary to support long term sustainability. For "implementing" PREMIS, the project did not specify how metadata was to be stored but recommended:

- use of PREMIS as a checklist against which repositories could compare their own preservation metadata
- inclusion of PREMIS in a profile for exchanging preservation metadata

METS was chosen for the profile because it was the best understood of the standards for exchanging metadata about digital objects, it was the standard being used and discussed in the PREMIS implementors' group, and it could be applied to different types of digital objects.

The profile provides a concrete framework in which to implement PREMIS. Repositories could demonstrate they met the preservation metadata requirements by being able to produce documents conforming with the profile.

From the original draft workplan the following tasks were carried forward:

- Identify the elements from the PREMIS Preservation Metadata Framework that would need to be mandatory in the Australian repository environment, taking into account the file formats most likely to be supported and scenarios for future use of this metadata for preservation management purposes.
- As part of this process distinguish elements that can be automatically generated from supported file formats as part of the ingest process.
- Assess the extent to which the selected repositories already support the collection of mandatory preservation metadata elements.
- Develop functional requirements for enhancing the selected repositories to support the collection of preservation metadata.
- Establish the profile for exchanging preservation metadata.

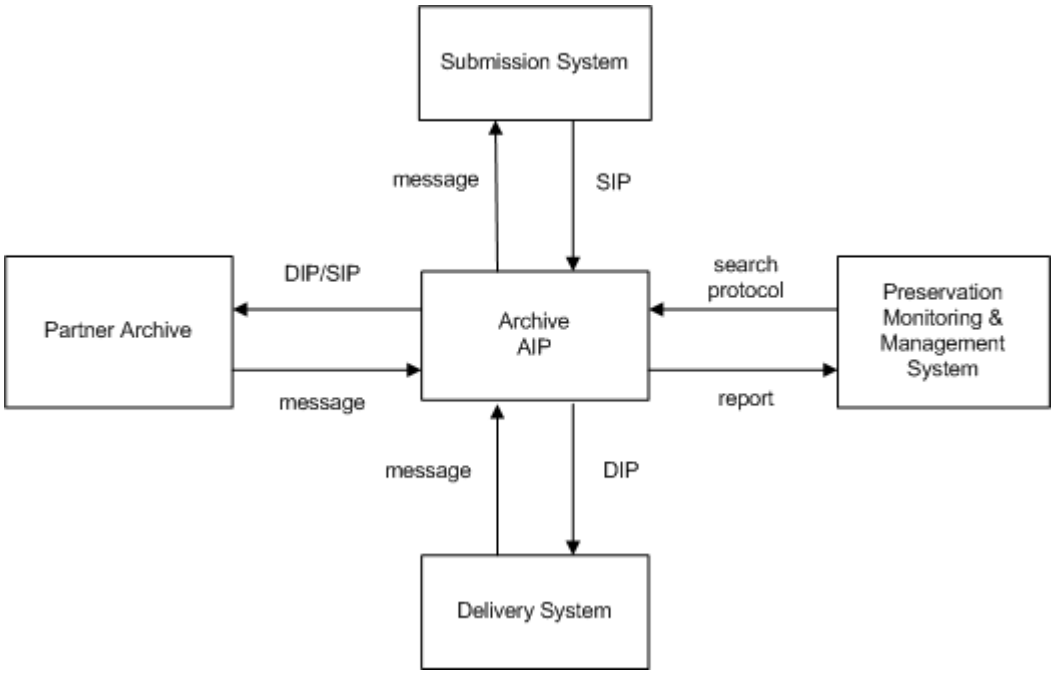
4. Products of the project

The main products of the project were:

- List of preservation metadata elements
- Supported file formats
- Tools for automated metadata collection
- Gap reports for ANU and UQ
- Preservation event use cases
- Profile for exchanging metadata

Other products were included as part of the original work plan. The products are in the appendices.

4.1. Service framework



This diagram shows part of an archive service framework (for more see for example the Fedora service framework (2005-2007). SIP (Submission Information Package), AIP (Archival Information Package), DIP (Dissemination Information Package) are concepts from Reference Model for an Open Archival Information System (OAIS).

The AIP in the **archive** should contain all the metadata required for long term sustainability and access. The AIP is conceptual: the metadata will not necessarily be stored as a single package in the repository and some metadata may be implicit (e.g. because it applies to all objects in the repository) rather than stored explicitly. The first product, the "List of preservation metadata elements", applies to the AIP.

The SIP is the object and metadata submitted, for instance by a depositor or harvester, through a **submission system**. It would have contained a subset of what is in the AIP. The product "Gap reports for ANU and UQ" look at what metadata is collected on submission and ingest.

The DIP containing an object and metadata sent to a **delivery system** for presentation to a user will also contain a subset of the AIP's metadata. This project didn't look at this area.

The **preservation monitoring and management system** may identify objects which need some preservation action through a search protocol and report and may then perform those actions. The product "Preservation event use cases" pertains to this area.

The archive may produce a DIP when transferring custody of an object to a **partner archive**. This becomes a SIP for ingest into the partner archive. The product "Profile for exchanging metadata" pertains to this area.

4.2 List of preservation metadata elements

This report in Appendix 1 details the metadata elements required for preservation purposes, i.e. metadata needed in order to provide meaningful long term access to digital objects. Metadata includes

- **PREMIS "core" preservation metadata**. The project specifically looked at Object, Event, Agent.

- **Descriptive metadata** (describes content, including metadata providing context or meaning to a digital object). Equivalent to "Intellectual Entity" in PREMIS. PREMIS doesn't go into detail about Intellectual Entity because it is covered by other schemas.
- **Structural metadata** - Needed for a repository to be able to reconstitute a whole digital object from its parts. A repository also needs to be able to display or present an object in a way that allows a user to understand how an object is related to is parts or to a greater whole. The current PREMIS relationship is not well suited to this.
- **Format specific metadata** e.g. image, audio, which was out of scope for PREMIS
- **Access rights metadata** - that is, metadata describing restrictions, permissions, conditions on use of an item which a repository must enforce or support when providing access to the item. PREMIS concentrated on permissions granted to the repository itself to carry out actions related to an item. PREMIS didn't examine access rights management in detail and neither did this project, as this is a complex and evolving area. The National Library of Australia is considering a rights management project for its digital and non-digital collections.

This report also includes a list of **"mandatory" elements**, that is, things a repository should know about *every* object. PREMIS and this report do not specify how metadata is to be stored or even if it is stored, perhaps because it applies to every object, for instance storageMedium. However if an element is not stored explicitly for each object it should be documented explicitly somewhere, e.g. in policy or procedures.

4.3 Recommended list of supported formats

This report in Appendix 2 is a "recommended list of supported formats". It is divided into material type, then for image, audio and video is further subdivided into recommended archival formats, formats in common usage which may be supported, e.g. formats produced by digital cameras and recording equipment, and unsupported formats. Most repositories will not support *all* of the recommended formats. File formats under "unsupported formats" and others *not* on this list should be converted to another format before being accepted by a repository because they are likely to be difficult to support in the long term.

Not included are specialist file formats which would be kept in specialist data repositories e.g. FITS (Flexible Image Transport System) used to manage astronomical data. Formats intended for delivery purposes, such as streaming media, are also not included.

4.4 Tools for automated metadata collection

This report in Appendix 3 recommends tools for identifying file formats and automatically extracting metadata. It includes an evaluation of the tools' capabilities and examples of output showing metadata that can be automatically generated. Output does not include empty elements for metadata not present in or not applicable to the files the tools are used on, and therefore the examples may not fully represent the capabilities of the tools. The National Library of Australia intends to do a more detailed audit to align metadata able to be output against recommended preservation metadata elements.

4.5 Gap reports for ANU DSpace and UQ Fez/Fedora repositories

Gap reports for ANU and UQ are in Appendix 4 of this report. The report assesses the extent to which the ANU and UQ repositories already support the collection of preservation metadata elements and includes recommendations for enhancements where gaps were identified. The most significant gaps were:

- recording of preservation events
- recording of structural relationships
- file format validation (ANU)
- checksum generation (UQ)

4.6 Preservation Event use cases

This document describes the requirements for actions that need to be taken on objects in a digital preservation repository and recording those actions or events. The following use cases are described:

- **Performing an action on an object which doesn't change the object** e.g. error checking.
- **Performing an action on an object which transforms an object into a new object (without materially changing its content)** e.g. migration to a newer format.
- **Deleting an object**
- **Updating the content of an object:** An action performed in some repositories, not usually for preservation purposes, but included for clarification.
- **Updating metadata about an object:** It is desirable from a preservation point of view to have the most complete, accurate metadata available, therefore there needs to be a way of updating the metadata as new information comes to light.

4.7 Profile for exchanging metadata

A draft METS profile is proposed in Appendix 8 in the form of a table of rules and recommendations. The National Library of Australia needs to test the profile with ANU and UQ and after further consultation with them and the wider digital preservation community, revise the profile. It can then expressed in xml using the formal METS profile schema and submitted to METS for registration. The scenario this profile addresses is transferring custody of an object from one repository to another because this is the scenario that requires the full set of preservation metadata. The draft profile is meant to be a common non-system specific profile which APSR partner repositories can map their system-specific requirements to.

5. Summary of recommendations from the project

These are the recommendations arising from the products above.

1. Repositories collect the full range of metadata necessary to provide meaningful long-term access to digital objects:
    - core preservation metadata (PREMIS)
    - descriptive metadata (describes content including metadata providing context or meaning to a digital object)
    - structural metadata (how parts relate to the whole and to each other)
    - file format specific metadata (e.g. image, audio formats)
    - access rights metadata (so material can be made available in accordance with rightsholders' conditions)
2. Repositories ensure they collect the mandatory PREMIS core preservation metadata elements in section A1.6.
3. Repositories aim to collect non-mandatory PREMIS metadata where applicable.
4. Repositories have policies and procedures which encourage deposit of digital material in open, standard formats.
5. Repositories have policies and procedures which articulate the level of support provided for particular formats.
6. Repositories identify and validate the file formats of objects on ingest or shortly thereafter.
7. Repositories use tools on ingest of an object or periodically on new objects, to collect extra metadata and/or metadata which can't be easily supplied during submission.
8. The National Library of Australia embark on a more detailed audit to align metadata able to be output against recommended preservation metadata elements and make results of this work available when ready.
9. The Australian National University (ANU) and the University of Queensland (UQ) repositories consider implementing the enhancements suggested in the gap reports, particularly
    - recording of preservation events
    - recording of structural relationships
    - file format validation (ANU)

    - checksum generation (UQ)

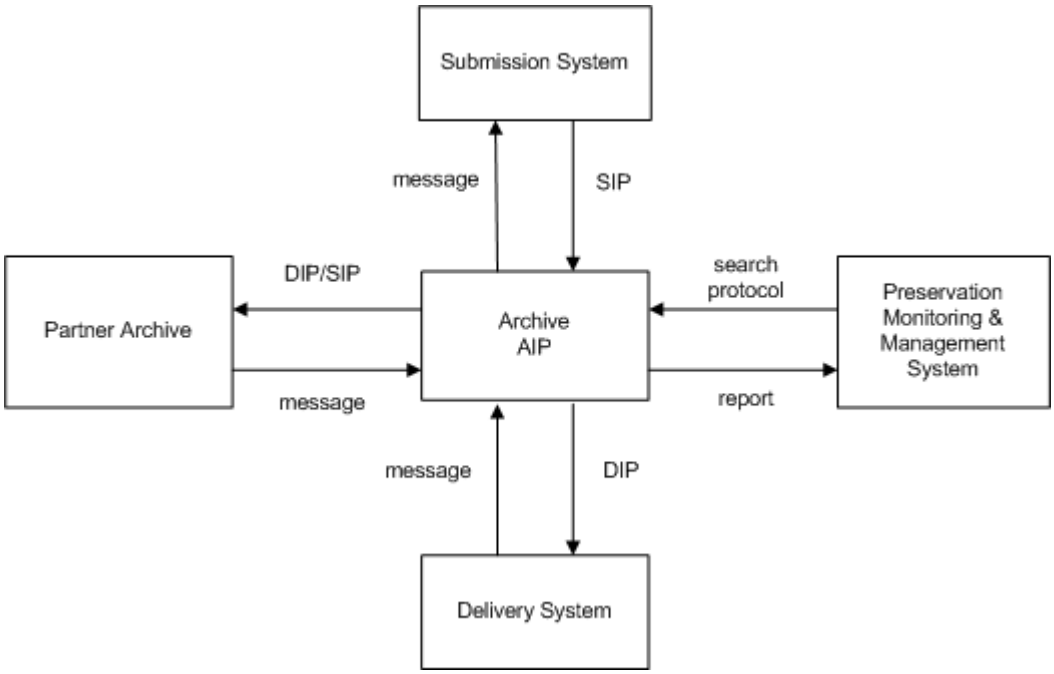
10. ANU and UQ repositories particularly take note of the functional requirements for preservation events in Appendix 6 and bring them to the attention of their open source communities as they begin to develop event logging functionality.
11. Australian repositories, particularly the National Library of Australia, continue to actively participate in development of standards relevant to digital preservation.
12. Australian repositories continue to actively participate in development of open source software for digital repositories, encouraging support for digital preservation metadata and standards in these developments.
13. The National Library of Australia develop crosswalks, if not already available, to map elements from schemas output by automated tools to PREMIS where an equivalent element exists.
14. The National Library of Australia continue to develop and test the proposed METS profile for metadata exchange with input from the Australian National University and the University of Queensland and consultation with the wider digital preservation community with a view to registering the profile formally.

Appendix 1: Preservation metadata elements

This part of the report details the metadata elements required for preservation purposes, i.e. metadata needed in order to provide meaningful long term access to digital objects. It also comments on elements which should be "mandatory".

A1.1 Service framework

The following diagram shows part of an archive service framework (for more see for example the Fedora service framework (2005-2007). SIP (Submission Information Package), AIP (Archival Information Package), DIP (Dissemination Information Package) are concepts from the Reference Model for an Open Archival Information System (OAIS).



This report is concerned with the AIP which should contain all the metadata required for long term sustainability and access. The SIPs (e.g. the object and metadata submitted by a depositor or harvester) and the DIPs (e.g. object and metadata sent to a system for presentation to a user) will contain subsets of the AIP's metadata. The AIP is conceptual: the metadata will not necessarily be stored as a single package in the repository and some metadata may be implicit (e.g. because it applies to all objects in the repository) rather than stored explicitly.

A1.2 Scenarios

To determine the requirements for preservation metadata, one needs to consider the uses to which the metadata will be put. Preservation metadata will be needed to support the following general scenarios:

- provide long term access - allow a digital object to be found and retrieved at some point in the future
- access, render, display or execute a digital object and allow the rendered content to be interpreted and understood by its intended users
- prove authenticity of an object including keeping a history of any changes to the object
- identify objects at risk in order to take some preservation action on them
- support repository planning and management e.g. to estimate resources (e.g. time, storage capacity) needed to undertake particular tasks on particular sets of objects

- to be able to restore or recreate a digital object e.g. an error may be discovered in a transformation process years after the event
- be able to transfer individual objects, or sections of or a whole archive, to another archive for safekeeping

While these general scenarios above seem straightforward, they could represent many different specific scenarios. For instance a repository receives an access request for an object, retrieves the object but then is unable to render it correctly. It is hard to imagine all the problems that might occur even 10 years years in the future, let alone over a longer period. Therefore it seems wise to collect as much metadata as possible "just in case", as the metadata may not be able to be obtained "just in time" when problems arise in the future.

A1.3 Metadata elements

The PREMIS Data Dictionary provides the "core preservation metadata element set". It was accepted that the elements were *all* necessary where applicable. It only remained to examine PREMIS to see how it should be interpreted and implemented.

However PREMIS's scope is deliberately limited to metadata which could apply to *all* digital objects regardless of format. It does not include file format specific metadata. It includes but does not go into detail about Intellectual Entity because "descriptive metadata is well served by existing standards". It also limits itself to "characteristics of rights and permissions concerned with preservation activities, not those associated with access and/or distribution".

Metadata needed for sustainable long-term access should therefore include not only PREMIS metadata but also:

- **PREMIS "core" preservation metadata**. The project specifically looked at Object, Event, Agent
- **Descriptive metadata** (describes content, including metadata providing context or meaning to a digital object). Equivalent to "Intellectual Entity" in PREMIS. PREMIS doesn't go into detail about Intellectual Entity because it is covered by other schemas
- **Structural metadata** - Needed for a repository to be able to reconstitute a whole digital object from its parts. A repository also needs to be able to display or present an object in a way that allows a user to understand how an object is related to its parts or to a greater whole.
- **Format specific metadata** e.g. image, audio, which was out of scope for PREMIS
- **Access rights metadata** - that is, metadata describing restrictions, permissions, conditions on use of an item which a repository must enforce or support when providing access to the item. PREMIS concentrated on permissions granted to the repository itself to carry out actions related to an item. PREMIS didn't examine access rights management in detail and neither did this project, as this is a complex and evolving area. The National Library of Australia has embarked on a rights management project for its digital and non-digital collections.

This report should be used by repositories as a checklist against which to compare their own preservation metadata specification. It does not specify how or even whether metadata elements should be stored. There is no expectation, for instance, that the PREMIS elements will be stored as a group or as a discrete set of metadata, although they could be. On the contrary, the elements are likely to be stored in various places and some will be implicit perhaps because they apply to every object in the repository. However it is important to note that metadata not stored explicitly for each object should be *documented explicitly* somewhere e.g. in repository policy or procedures.

A repository can demonstrate its ability to meet these requirements by producing a document conforming to the draft METS profile in Appendix 8.

A1.4 What does "mandatory" mean?

In terms of this Appendix, "mandatory" means a piece of metadata a repository is expected to "know" about each object to which the metadata applies, whether the metadata is stored explicitly or not.

In the draft METS profile for metadata exchange in Appendix 8, "mandatory" means the element must be present in a conforming METS document. More mandatory elements are specified in the METS profile as stricter requirements aid system interoperability. If data is unable to be supplied for a mandatory element, the element may contain values "not_applicable" or "unknown".

A1.5 Core preservation metadata (PREMIS elements)

The elements in the PREMIS Data Dictionary are "core preservation metadata" elements. ALL elements should be collected by the repository if applicable. The elements are listed below with some brief notes on applying them. Object entity elements apply to the objectCategory "file".

See the PREMIS Data Dictionary for fuller definitions, rationale, obligation (i.e. mandatory or optional), repeatability, usage and other notes.

Comments have been made against some elements on how to apply them in the draft APSR METS profile.

A1.5.1 Object entity elements

- **objectIdentifier**

  The object must be uniquely identified within the repository in which it is stored, so this is mandatory.

  - **objectIdentifierType**
  - **objectIdentifierValue**
- **preservationLevel**

  The draft METS profile proposes four values: "supported", "known", "unsupported" and "not_applicable". "supported" means the format is fully supported (the repository is confident of maintaining accessibility of the object in the long term); "known" means not fully supported yet but the object has a high priority for continued access (the repository will attempt to obtain enough information to enable the format to be upgraded to the "supported" level); "unsupported" means not fully supported and low priority (the bitstream will be preserved as is but no action will be taken to guarantee continued access); "not_applicable" means there is no intention to preserve an object (e.g. there may be another representation of the item which is the preservation copy or this may be an interim draft of a document). More detailed information about preservation intentions can be stored in notes fields in the descriptive metadata.

- **objectCategory**

  Suggested values are "representation", "file" and "bitstream" (as defined by PREMIS).

- **objectCharacteristics**
  - **compositionLevel**

    If, as is the usual case, there is only one composition level (e.g. no decompression or decryption needed to recover the original object), this defaults to "0".

  - **fixity**

It is expected that every repository keep information to verify that a file hasn't changed. It is recommended that this be mandatory, even though it is optional in PREMIS.

- **messageDigestAlgorithm**
- **messageDigest**
- **messageDigestOriginator**

  Default to the repository name unless the messageDigest originated elsewhere. It is expected that a repository calculate its own checksums. If the file submitted to the repository already has a checksum, we would expect the repository to verify it. Therefore the value may not be the repository name in a SIP but should always be the repository name in an AIP.

- **size**

It is expected that a repository to be able to determine this and it has a number of uses for delivery and preservation management, so it is recommend that this be mandatory, even though it is optional in PREMIS.

- **format**
  - **formatDesignation**

    It is recommended that formatDesignation should be used in addition to formatRegistry because a) the formatRegistryKey may not be as informative as formatName and formatVersion; b) the format registry may not be available when needed and the repository has no control over it; and c) it may be needed for preservation management searching and reporting e.g. retrieving all PDFs, regardless of version

    - **formatName**
    - **formatVersion**
  - **formatRegistry**

    Prefer universally available and more complete registries e.g. GDFR and PRONOM, which are under development. More than one registry entry may be cited.

    - **formatRegistryName**
    - **formatRegistryKey**
    - **formatRegistryRole**
- **significantProperties**

A place to record important characteristics which can't be recorded anywhere else. Where this applies, it is a measure of preservation success. The value of this element is a matter of subjective human judgement. This is an unstructured element in PREMIS but it may be possible to develop structured descriptions of significant properties based on an object's class or content (in a similar way that complex rights can be described in a structured way in XACML), which, in conjunction with repository policies, may either assist automated population of this element or assist automation of actions to be taken on objects in response to this element.

- **inhibitors**

Used for e.g. encryption, passwords. Will only apply to some objects.

- **inhibitorType**
- **inhibitorTarget**

        - **inhibitorKey**
- **creatingApplication**

  Can be useful for problem solving purposes e.g. it is not uncommon for certain versions of software to be known for causing conversion errors or introducing artifacts.

    - **creatingApplicationName**
    - **creatingApplicationVersion**
    - **dateCreatedByApplication**

      Many repositories only store the date an object was ingested, not the date it was created by the application. The date created by the application is part of the object's provenance and may also be useful for problem solving purposes.

- **originalName**

  It is important to be able to identify the file by its original name, e.g. to communicate with depositors who may only know the file by its original name). It may also be needed to reconstruct internal links. It should be mandatory even if the original name is the same as the repository filename.

- **storage**
    - **contentLocation**
        - **contentLocationType**
        - **contentLocationValue**
    - **storageMedium**

      The repository needs to know the medium on which an object is stored in order to know how and when to do media refreshment and media migration. In some cases the value may not be the specific medium but the storage system that knows the medium. It may not be explicit for each object but should be centrally recorded explicitly for the repository.

- **environment**

  Environment (hardware/software combinations supporting use of the object) doesn't need to be mandatory because it shouldn't be critical for recommended archival formats and the information may not be available. It would be preferable to refer to a registry of environment information rather than store environment for individual objects, except for special cases. Such a registry would ideally record the kinds of systems that were available at particular points in time, in addition to describing environments for particular file formats, because environment can't always be inferred from the format or filename extension (e.g. ".exe" files). If the institution has a standard environment and the file runs on that, the repository could note it. For special cases the deposit workflow could automatically supply details of the machine of the depositor (e.g. from a local registry).

    - **environmentCharacteristic**

      PREMIS suggests values: unspecified, known to work, minimum, recommended. For obscure formats and complex objects not covered by registries, submitters/repositories should be encouraged to provide a "known to work" environment.

    - **environmentPurpose**
    - **environmentNote**
    - **dependency**
        - **dependencyName**

This should be used in addition to dependencyIdentifier, as it may not be self-evident from the dependencyIdentifier what the nature of the dependency is.

        - **dependencyIdentifier**
            - **dependencyIdentifierType**
            - **dependencyIdentifierValue**
    - **software**
        - **swName**
        - **swVersion**
        - **swType**
        - **swOtherInformation**
        - **swDependency**
    - **hardware**
        - **hwName**
        - **hwType**
        - **hwOtherInformation**
- **signatureInformation**

  Information needed to use a digital signature to authenticate the signer of an object and/or the information contained in the object. Will only apply to some objects.

    - **signatureInformationEncoding**
    - **signer**
    - **signatureMethod**
    - **signatureValue**
    - **signatureValidationRules**
    - **signatureProperties**
    - **keyInformation**
        - **keyType**
        - **keyValue**
        - **keyVerificationInformation**
- **relationship**
    - **relationshipType**

      PREMIS notes imply this element can be used to describe relationships of parts to a whole (structural context) as well as to describe how one object has been derived from another (provenance). A repository should record all significant relationships. However in the draft APSR METS profile, structural relationships will be recorded in the METS element structMap, not in a PREMIS relationship element. PREMIS relationship will be reserved for provenance relationships.

    - **relationshipSubType**

      Note that the relationship should be described from the point of view of the current object entity (eg the current object "is a derivative of" the related object).

    - **relatedObjectIdentification**
        - **relatedObjectIdentifierType**
        - **relatedObjectIdentifierValue**
        - **relatedObjectSequence**
    - **relatedEventIdentification**

      An event associated with the relationship. Only applies to derivation relationships. Structural relationships won't usually have associated events.

        - **relatedEventIdentifierType**
        - **relatedEventIdentifierValue**
        - **relatedEventSequence**
- **linkingEventIdentifier**

Use to link to events that are not associated with relationships between objects, such as format validation, virus checking etc.

    - **linkingEventIdentifierType**
    - **linkingEventIdentifierValue**
- **linkingIntellectualEntityIdentifier**

  This may be a link to descriptive metadata that describes the Intellectual Entity. this link may be to an identifier of an object that is at a higher conceptual level than the object for which the metadata is provided, e.g. to a collection or parent object.

    - **linkingIntellectualEntityIdentifierType**
    - **linkingIntellectualEntityIdentifierValue**
- **linkingPermissionStatementIdentifier**
    - **linkingPermissionStatementIdentifierType**
    - **linkingPermissionStatementIdentifierValue**

A1.5.2 Event entity elements

For more information on events, see the event use cases in Appendix 6.

- **eventIdentifier**
    - **eventIdentifierType**
    - **eventIdentifierValue**
- **eventType**
- **eventDateTime**
- **eventDetail**
- **eventOutcomeInformation**
    - **eventOutcome**
    - **eventOutcomeDetail**
- **linkingAgentIdentifier**
    - **linkingAgentIdentifierType**
    - **linkingAgentIdentifierValue**
    - **linkingAgentRole**
- **linkingObjectIdentifier**

  Any change to an object creates a new object. Although repeatable, there isn't an element to designate which is the source object and which is the new object. However this information can be determined from the Object Entity relationship element.

    - **linkingObjectIdentifierType**
    - **linkingObjectIdentifierValue**

      While considering the linking of Ojbects to Events, there was discussion of a fixity check Event where the outcome was failure, i.e. the checksum is not the same as it was before, indicating the file has been corrupted or changed. The changed object may be replaced with a new object (with a new objectIdentifier) derived from the original but if that were not possible and the changed object were usable, should it be regarded as the same or a different object?

A1.5.3 Agent semantic units

Agents may be persons, organisations, or software, associated with rights management and preservation events in the life of a data object.

- **agentIdentifier**
    - **agentIdentifierType**
    - **agentIdentifierValue**
- **agentName**

- **agentType**

A1.5.4 Rights semantic units

- **permissionStatement**

  An agreement with a rights holder that allows a repository to take action(s) related to objects in the repository.

  - **permissionStatementIdentifier**
    - **permissionStatementType**
    - **permissionStatementValue**
  - **linkingObject**
  - **grantingAgent**
  - **grantingAgreement**
    - **grantingAgreementIdentification**
    - **grantingAgreementInformation**
  - **permissionGranted**
    - **act**
    - **restriction**
    - **termOfGrant**
      - **startDate**
      - **endDate**
    - **permissionNote**

A1.6 Mandatory PREMIS elements

This is the list of PREMIS elements mandatory for APSR repositories. It is a checklist of things a repository should know about EVERY object in the repository. If the information is not recorded explicitly about each object, it should be able to be determined from the repository itself or from repository documentation of policies, procedures etc. The draft METS profile in Appendix 8 also specifies mandatory elements for conforming METS documents.

A1.6.1 Object Entity elements:

The following elements are mandatory in the PREMIS data dictionary for objectCategory "file". They are not necessarily mandatory in the PREMIS xml schema since they may not apply to all types of objectCategory.

- objectIdentifierType
- objectIdentifierValue
- preservationLevel
- objectCategory
- compositionLevel
- storageMedium

The following additional elements from PREMIS were regarded as important enough to be mandatory for APSR repositories by the project working group.

- messageDigestAlgorithm
- messageDigest
- size
- formatName
- originalName

A1.6.2 Event Entity elements:

PREMIS does not mandate the existence of an Event. An Event can be linked to an Object through the Object entity's optional relationship or linkingEventIdentifier elements, or it can be linked to an Object through the Event's optional linkingObjectIdentifier element.

However we recommend the following be mandatory:

- knowledge of an Ingest Event is mandatory for every object (date of Ingest is when the Object is actually stored in the repository)
- any event which changes an Object must always be recorded

Validation events should be recorded e.g. that a file is of the format it says it is. Where validation is done on every file on ingest or a validation tool is run over a whole repository at a particular point in time, this fact should be recorded by the repository. Events are examined in more detail in the Event use cases in Appendix 6.

The mandatory elements (i.e. things the repository must know about every event) are:

- eventIdentifierType
- eventIdentifierValue
- eventType
- eventDateTime

The Event should "know" about the Object/s it acted on. However PREMIS does not specify a mandatory link between Events and Objects either in the Object entity or the Event entity. In the APSR draft METS profile, it will be mandatory for the Object entity to contain a linkingEventIdentifier to the (mandatory) Ingest event, and the Event entity will not need to contain the reciprocal linkingObjectIdentifier.

A1.6.3 Agent

PREMIS does not mandate the existence of an Agent entity since linkingAgentIdentifier is optional in the Event entity.

However we recommend that repositories should know about Agent if the Event is one which changes an Object. Agent should, for example, identify the software used. Although the software may already be described in the Object entity (in creatingApplication) or in the Event entity, placing it in Agent in a document conforming with the APSR METS profile will facilitate mapping in the receiving repository's database. Agent should also be used for an organisation if an organisation *other* than the transferring repository was responsible for an Event.

A1.6.4 Rights entity

PREMIS does not mandate the existence of Rights since linkingPermissionStatementIdentifier is optional in the Object entity. PREMIS concentrated on rights concerned with preservation activities.

Rights should be mandatory in so far as repositories should have agreements, or some conditions which depositors agree to when they deposit material in the repository, in place, but this may not apply when material is out of copyright.

A1.7 Descriptive metadata

Descriptive metadata is considered mandatory for APSR repositories but this project did not examine this area in detail and does not prescribe a particular metadata scheme. Repositories will need to be able to output descriptive metadata in MODS to conform with the APSR METS profile

for metadata exchange, but should store and be able to output descriptive metadata in a form which retains the granularity of all available metadata.

Descriptive metadata includes not only metadata such as creator, title, date, subjects, but also contextual metadata. Contextual metadata provides meaning to or aids interpretation of an object.

A1.8 Structural metadata

For some objects structural metadata is needed for a repository to be able to reconstitute a whole digital object from its parts. A repository also needs to be able to display or present an object in a way that allows a user to understand how an object is related to its parts or to a greater whole. It may be stored in a PREMIS Object entity "relationship" element but may be better stored as a structural map, a manifest of files or a set of relationships.

A1.9 File format specific metadata

File format specific metadata is needed to record the characteristics of a digital object so that it can be accurately rendered. In some cases, without file format specific metadata a system may not be able to render a digital object at all.

The following metadata schemes and extensions to them proposed by this project are recommended for use in the APSR METS profile. The schemas may include mandatory elements. The extensions will be published on the National Library of Australia website.

This section is intended to be used as a supplement to the Library of Congress Audiovisual Prototyping Project, http://www.loc.gov/rr/mopic/avprot/(2004). Indicated in this section are alternative field names which have been used in the National Library of Australia's Digital Collections Manager (DCM) as well as a set of additional metadata fields which is itself an extension to the Library of Congress METS extension schema.

It is likely that automated harvesting of data for many of these metadata fields is currently not be possible, however recording such data will assist in long-term management of the files.

It should also be noted that the set of suggested additional metadata fields is not necessarily complete and it is intended that other organisations/institutions provide further input.

A1.9.1 Image

MIX is the recommended extension schema to be used for image metadata. MIX is a schema endorsed by the METS Editorial Board for use with METS. The following is the introduction from the MIX home page:

The Library of Congress' Network Development and MARC Standards Office, in partnership with the NISO Technical Metadata for Digital Still Images Standards Committee and other interested experts, is developing an XML schema for a set of technical data elements required to manage digital image collections. The schema provides a format for interchange and/or storage of the data specified in the NISO Draft Standard Data Dictionary: Technical Metadata for Digital Still Images (Version 1.2). This schema is currently in draft status and is being referred to as "NISO Metadata for Images in XML (NISO MIX)". MIX is expressed using the XML schema language of the World Wide Web Consortium. MIX is maintained for NISO by the Network Development and MARC Standards Office of the Library of Congress with input from users.

MIX is the recommended extension schema to be used for image metadata.

MIX schema

A1.9.2 Audio and Video

A1.9.2.1 Audio

The Library of Congress Audio (Source) Data Dictionary, which was developed as part of the Audio-Visual Prototyping Project, is the recommended base extension schema for audio metadata. It can be found at http://www.loc.gov/rr/mopic/avprot/DD_ASMD.html. The following metadata fields are intended as a further extension to the Library of Congress Audio (Source) Data Dictionary.

file_format
: The type of audio file for any audio file, for example a WAV file is a Microsoft WAVE file and an AIF is the Audio Interchange File Format.

file_version
: The version of the file format used.

coding_history
: Indicates the file format history (and devices) that the file has been through.

mime_type
: The MIME type helps web browsers associate particular files with suitable player applications or plug-ins.

compression
: The type of compression used on the file – for "non-archival" quality files where an archival copy is not available – this may be something such as MPEG compression such as in an MPEG 1 Layer 3 file (MP3).

codec_version
: The version of the codec used (if appropriate).

file_container
: The type of file format that is used to hold another file format, for example the Broadcast Wave Format (BWF) is a file container for a Microsoft WAV file.

file_container_version
: The version of the file container format used.

frame_rate
: The number of frames per second.

byte_order
: For example "Big Endian" or "Litte Endian".

timecode_type
: Type of time code recorded on the audio source item, for example: SMPTE drop frame, SMPTE non drop frame, etc.

channel_num
: Indicates the specific channel number, for example channel number 0. This is a repeatable field.

channel_num_map_loc
: This is tied to the specific channel number and should indicate the position of channel, for example, channel number 0 in a stereo file for the channel number map location may indicate "left".

channel_map_config
: The configuration of the mapping of channels. This information is important for multichannel works. Examples of configuration are the "shoebox" and "double diamond".

delivery_type
: For streaming media files (this, for instance, could indicate RTSP). While streaming media files are not recommended for inclusion as they are of a non-archival format, in the instance that an exception is made to include streaming media, it is necessary for a record of intended delivery protocol to be available. For example: QuickTime files that are produced with the settings, “hinted for streaming” indicate that these files can only be accessed using the RTSP protocol.

encoding_software
: Software used to encode the software for delivery files (only necessary in the exception of non-archival quality delivery files).

codec_essence
: The particular type or “flavour” of the codec used for example RealMedia “Music” or “Voice” codec.

codec_essence_version

The version of the codec essence used.

A1.9.2.2 Video

The Library of Congress Video (Source) Data Dictionary, which was developed as part of the Audio-Visual Prototyping Project, is the recommended base extension schema for video metadata. It can be found at http://www.loc.gov/rr/mopic/avprot/DD_VSMD.html. The following metadata fields are intended as a further extension to the Library of Congress Video (Source) Data Dictionary.

file_format
The type of video file, for example a MOV file is a QuickTime file format, however it should be noted that with video there is quite often both a file format and a container format.

file_version
The version of the file format used.

coding_history
Indicates the file format history (and devices) that the file has been through.

mime_type
The MIME type helps web browsers associate particular files with suitable player applications or plug-ins.

compression
The type of compression used on the file – for “non-archival” quality files where an archival copy is not available – this may be something such as MPEG compression such as in an MPEG 2, which is the format used for DVD presentation. While archival materials should not be stored in a compressed format, it should be noted that at the current period in time, video files are very large and due to other restraints (such as cost of large scale data storage infrastructures) storing uncompressed video is currently not always possible. Video is still a relatively unexplored field in relation to archiving and preservation and over time it is assumed that practices and standards for video archiving will change.

codec_version
The version of the codec used (if appropriate).

file_container
The type of file format that is used to hold another file format, for example the QuickTime (MOV) is a file container for other files formats. There can be similarities between the file format and file container formats.

file_container_version
The version of the file container format used.

byte_order
For example “Big Endian” or “Little Endian”.

counting_mode
NTSC drop-frame or non-drop frame.

track_num
Indicates the specific track number, for example channel number 0. This is a repeatable field. (This is similar to the audio metadata field channel_num.)

track_num_map_loc
This is tied to the specific track number and should indicate the position.

track_map_config
The configuration of the mapping of tracks. This information is important for (rare) works where more than one video track is present.

delivery_type
For streaming media files (this, for instance, could indicate RTSP). While streaming media files are not recommended for inclusion as they are of a non-archival format, in the instance that an exception is made to include streaming media, it is necessary for a record of intended delivery protocol to be available). For example: QuickTime files that are produced with the settings, “hinted for streaming” indicate that these files can only be accessed using the RTSP protocol.

encoding_software
Software used to encode the software for delivery files (only necessary in the exception of non-archival quality delivery files).

broadcast_standard
This includes PAL, NTSC, SECAM, DV, HDV etc

anamorphic

A playback presentation setting related to how DVD video has been mastered and whether the video is capable of being played back on screens with different aspect ratios. For example, this would include being able to play the video material on screens with either 4:3 and 16:9 aspect ratios without the video being “squashed” to fit. Values for this metadata field should either be “true” or “false”.

field_dominance
This is set to either lower (even) or upper (odd).

alpha_channel
Whether or not the video has an alpha_channel.

codec_essence
The particular type or “flavour” of the codec used for example RealMedia “Music” or “Voice” codec.

codec_essence_version
The version of the codec essence used.

A1.9.4 Text, HTML and XML

Schema for Technical Metadata for Text (created by Jerome McDonough, Elmer Bobst Library, New York University) is endorsed by the METS Editorial Board for use with METS.
Schema
Documentation

Further analysis of additional metadata fields required for text documents should be carried out.

A1.9.4.1 Additional metadata fields

markup_nature
Whether the mark-up style is strict or transitional (in the case of HTML and XHTML)

A1.9.5 Alternative naming

Terminology can vary. The following list of alternative names is provided for clarity.

audio_block_size
block align

audio_data_encoding
encoding

bits_per_sample
bit depth

codec_name
codec

num_channel
channels

sound_field
recording mode

file_container
wrapper or container

sampling_frequency
sampling rate

data_rate
bit rate

timecode_type
display format

pixels_horizontal
image width

pixels_vertical
image height

charset
encoding

1.10 Access rights metadata

It should be mandatory to record access rights for materials with restricted access conditions. If no access rights are recorded it would be assumed that there are no access restrictions.

This report does not prescribe a particular metadata scheme. Possibilities include METS Rights, PREMIS Rights, Creative Commons licences and XACML.

1.11 Recommendations

1. Repositories collect the full range of metadata necessary to provide meaningful long-term access to digital objects:
    - core preservation metadata (PREMIS)
    - descriptive metadata (describes content including metadata providing context or meaning to a digital object)
    - structural metadata (how parts relate to the whole and to each other)
    - file format specific metadata (e.g. image, audio formats)
    - access rights metadata (so material can be made available in accordance with rightsholders' conditions)
2. Repositories ensure they collect the mandatory PREMIS core preservation metadata elements in section A1.6.
3. Repositories aim to collect non-mandatory PREMIS metadata where applicable.

Appendix 2: Recommended list of supported formats

Appendix 2 comprises a list of formats likely to be supported by repositories. Most repositories will not support *all* of these formats. File formats *not* on this list are likely to be more difficult to support in the long-term. It is acknowledged that constant Information Technology development will produce new and improved archival formats, and it is intended that any new additions to this list be included where appropriate. The list not only includes recommended archival formats but also other formats likely to be accepted by repositories e.g. formats produced by digital cameras or recording equipment.

Archival formats should ideally be based on open standards, but widely used and supported, well documented proprietary formats may be acceptable. It should be noted that while appropriate archival and "commonly in use" formats have been listed here - this document does not indicate recommended quality standards for digital media items, and appropriate guidelines for such should be sought. Files containing any form of compression should be carefully considered.

Not included are specialist file formats which would be kept in specialist data repositories e.g. FITS (Flexible Image Transport System) used to manage astronomical data. Formats intended for delivery purposes, such as streaming media, particularly where formats are non-stand-alone and and dependent on specific protocols for access (such as RTSP), are also not included.

This list was developed in consultation with Kevin Bradley who co-authored Survey of data collections: a research project undertaken for the Australian Partnership for Sustainable Repositories.

A2.1 Images

2.1.1 Recommended Archival Formats

These formats are recommended archival formats and are included here in order of preference of the preferred archival format.

1. Tagged Image File Format (TIFF)

While TIFF is the recommended archival format, both Multi-part TIFF files and Multi-layered TIFF files are not necessarily considered archival file formats and where possible Multi-part TIFF files should be stored as sets of single images and Multi-layered TIFF files should be flattened to single layer images. Each repository may decide to develop their own policies regarding these variations of the TIFF format.

2.1.2 Formats in Common Usage

These formats are not recommended as archival formats, however they are in common usage and so are included. As they are not archival formats no order of preference is indicated. Files in the following formats should preferably have a copy created in an archival format where possible.

- JPEG2000 - this is a newly emerging lossless compression format, however implementation of the standard has been relatively slow, and the majority of software in common usage is currently unable to read this image file format.
- Digital Negative Format (DNG)
- Scalable Vector Graphics (SVG)
- Encapsulated PostScript (EPS)
- Portable Network Graphic (PNG)
- JPEG/JIFF Image (JPEG) - this is a lossy compression file format, not an archival format, however it is in wide usage and repositories may need to support this

2.1.3 Unsupported Formats

These formats are not archival formats and are not recommended as supported formats by repositories. Files in these formats should be converted to recommended archival formats before being accepted by a repository as they are likely to be difficult to support in the long-term.

- Graphic Interchange Format (GIF) - although not recommended as an archival format for general images, GIF is likely to appear in archived websites
- Photoshop Format (PSD)
- Windows OS/2 Bitmap Graphics (BMP)
- Macintosh Quickdraw/PICT Drawing (PICT)

A2.2 Audio

It should be noted that some audio formats are a combination of a container or "wrapper" format and a file content format, and so are essentially a combination of two formats.

2.2.1 Recommended Archival Formats

These formats are recommended archival formats and are included here in order of preference of the preferred archival format. It should be noted that with some AV formats they are a combination of a wrapper format as well as a file content format, and so are essentially a combination of two formats.

1. Broadcast Wave Format (BWF) - wrapper that contains the WAV file format. The wrapper holds additional metadata
2. Waveform Audio (WAV)
3. Audio Interchange File (AIFF)

2.2.2 Formats in Common Usage

These formats are not recommended as archival formats, however they are in common usage and so are included. As they are not archival formats no order of preference is indicated. Files in the following formats should preferably have a copy created in an archival format where possible.

- MP3 - this is a lossy compression file format, not an archival format, however it is in wide usage and repositories may need to support this
- MPEG-4 - this is a lossy compression file format, not an archival format, however it is becoming a more widely used format and repositories may want to consider supporting this. This is a content format that may be contained within another wrapper format, for example QuickTime MOV or MP4 file formats (utilising the MPEG-4 AAC codec)

2.2.3 Unsupported Formats

These formats are not archival formats and are not recommended as supported formats for repositories. Files in these formats should be converted to recommended archival formats before being accepted by a repository as they are likely to be difficult to support in the long-term.

- Ogg Vorbis Codec Compressed WAV File (OGG) - open standard, however it is in limited public use
- Real Media (RM) - proprietary lossy compression streaming media file
- Windows Media File (WMV) - proprietary lossy compression streaming media file

A2.3 Video

These formats are recommended archival formats and are included here in order of preference of the preferred archival format. It should be noted that with some AV formats they are a combination

of a wrapper format as well as a file content format, and so are essentially a combination of two formats.

2.3.1 Recommended Archival Formats

Currently there is no archival video standard, however a number of options are available. Unlike other media types such as audio or image, video requires large amounts of storage space. For this reason, some compressed formats are currently considered to be suitable (for the time being) as archival formats until storage of large video files plus recommended archival video standard becomes a reality. These formats are recommended archival formats and are included here in order of preference of the preferred archival format.

1. Material Exchange Format (MXF) - wrapper that contains a range of "essence" or "content" file formats. The wrapper holds additional metadata
2. Advanced Authoring Format (AAF) - wrapper that contains a range of "essence" or "content" file formats. The wrapper holds additional metadata
3. MOTION JPEG2000 (MJ2) - this is a newly emerging lossless compression format, however implementation of the standard has been relatively slow, and the majority of software in common usage is currently unable to read this image file
4. MPEG-2 - lossy compression format. This is the standard used for DVD

2.3.2 Formats in Common Usage

These formats are not recommended as archival formats, however they are in common usage and so are included. As they are not archival formats no order of preference is indicated. Files in the following formats should preferably have a copy created in an archival format where possible.

- Digital Video Digital Interface Format (DV-DIF) - Raw file format for digital video
- MPEG-1 - Moving Picture Experts Group early AV standard
- QuickTime Movie (MOV)
- MPEG-4 - this is a lossy compression file format, not an archival format, however it is becoming a more widely used format and repositories may want to consider supporting this. This is a content format that may be contained within another wrapper format, for example QuickTime MOV or MP4 file formats (utilising the MPEG-4 AAC codec)
- Audio Video Interleave (AVI)

2.3.3 Unsupported Formats

These formats are not archival formats and are not recommended as supported formats for repositories. Files in these formats should be converted to recommended archival formats before being accepted by a repository as they are likely to be difficult to support in the long-term.

- Real Media (RM) - proprietary lossy compression streaming media file
- Windows Media File (WMV) - proprietary lossy compression streaming media file

A2.4 Text

2.4.1 Recommended Archival Formats

These formats are recommended archival formats and are included here in order of preference of the preferred archival format. While formats such as Microsoft Word are commonplace, it should be noted that this is a proprietary format and is likely to be difficult to support in the long-term.

1. Extensible Markup Language (XML)
2. American Standard Code for Information Interchange (ASCII) Text (TXT)
3. 8-bit Unicode Transformation Format (UTF-8) Text (TXT)
4. 16-bit Unicode Transformation Format (UTF-16) Text (TXT)

2.4.2 Formats in Common Usage

These formats are not recommended as archival formats, however they are in common usage and so are included. As they are not archival formats no order of preference is indicated. Files in the following formats should preferably have a copy created in an archival format where possible.

- Open Document Format (ODF)
- Rich Text Format (RTF)

2.4.3 Unsupported Formats

These formats are not archival formats and are not recommended as supported formats by repositories. Files in these formats should be converted to recommended archival formats before being accepted by a repository as they are likely to be difficult to support in the long-term. However, it should be noted that some companies creating proprietary formats are considering developing future open format versions.

- Microsoft Word (DOC)
- Standard Generalized Markup Language (SGML)

A2.5 Databases

Databases contain a larger degree of complexity than other individual files. While a full analysis of database formats was not carried out, only databases with a simple structure are able to be supported. Databases containing complex relationships cannot be supported at this stage. In general, documentation of databases including rules and relationships should also be archived.

2.5.1 Recommended Archival Formats

These formats are recommended archival formats and are included here in order of preference of the preferred archival format. Only simple databases whose raw data can be turned into structured text, such as databases where all data can be extracted via a single join query, are considered a recommended archival format.

1. Extensible Markup Language (XML) - simple databases only
2. Comma-Separated Variables (CSV) - simple databases only

2.5.2 Formats in Common Usage

These formats are not recommended as archival formats, however they are in common usage and so are included. As they are not archival formats no order of preference is indicated. Files in the following formats should preferably have a copy created in an archival format where possible as proprietary formats are likely to be difficult to support in the long-term.

- Microsoft Access (MDB) - simple databases only
- Microsoft XL (XLS) - simple databases only

2.5.3 Out of Scope Formats

Complex databases were considered out-of-scope for this project and so are considered to be unsupported formats.

- Complex databases of any format
- Spreadsheets with macros

A2.6 Portable Document Format (PDF)

While the Portable Document Format (PDF) is a proprietary format, and proprietary formats are normally considered to be unsupported formats, PDF should currently be the exception. This is largely because it is a format in common usage and the large degree of academic papers are published and distributed in this format. Further work would need to be done on this format as it contains both text and image, and because there are several types of PDF, including PDF/A, a proposed archival standard for PDF accepted as an ISO standard in 2005.

A2.7 Websites

This project did not address websites specifically. The National Library of Australia is part of the International Internet Preservation Consortium which among other things is fostering the development of common tools, techniques and standards for website archiving

A2.8 Multimedia

Multimedia files (such as Director, Flash and Microsoft Powerpoint) were considered out-of-scope for this project. However, example output files from metadata extraction tools have been provided for a range of multimedia formats.

A2.9 Other Objects and Formats

Other formats that were considered out-of-scope of this project are considered unsupported formats currently.

- Learning objects - these are often a group of files with an xml manifest. There is no policy for these yet.
- Complex objects - not addressed in detail in this project

A2.8 Recommendations

1. Repositories have policies and procedures which encourage deposit of digital material in open, standard formats.
2. Repositories have policies and procedures which articulate the level of support provided for particular formats.

Appendix 3: Tools for automated metadata collection

A3.1 Introduction

Recommendations on the range of metadata elements to be collected by repositories are set out in Appendix 1 of this report. The degree to which repositories can meet such recommendations will depend on the metadata that can be re-used from existing records, policies and documentation, supplied by depositors, recorded as part of repository processes, or extracted from the materials themselves.

Given the volume of metadata that may be required or available, automated processes for collection of metadata are preferable, especially for metadata extraction from the materials themselves. A number of tools are available to address these needs in varying degrees and to provide some of the details required in an automated way. A selection of such tools are briefly described and compared in this Appendix. A more detailed alignment of metadata output from these tools against element recommendations will be made available when completed.

There are several aspects of metadata collection and the archiving process that may be addressed by tools:

- **File identification:** identifying file formats conclusively
- **Validation:** verifying that a file format is valid with respect to its specification
- **Generic metadata collection:** characterisation of files at a generic level
- **File format specific metadata collection**

At present, tools tend to cover one or more aspects of the archiving process and metadata collection, but no one tool yet covers all. Tools may also cover these aspects to varying degrees.

The range of formats covered by tools can also vary, and it may be useful to divide available tools into several classes, based on their format coverage:

- **Tools which handle a range of material types** (e.g. image, audio, text) and file formats. At present these are generally limited to identifying and characterising a small range of common formats, rather than a wide range of arbitrary formats. However, these are often also modular and extensible, so that other initiatives can create modules for characterising formats to suit their own needs.
    - Examples: DROID, JHOVE, National Library of New Zealand Metadata Extraction Tool (NLNZ-MET).
- **Tools which handle a single material type** (e.g. image), but which can deal with multiple file formats of that type.
    - Examples: ImageMagick (http://www.imagemagick.org/)
- **Tools which are limited to a single specific format.**
    - Examples: Readers or viewers for TIFF tags.

For the range of formats intended to be supported in APSR repositories, several tools may be suitable. It is likely that more than one will be needed to obtain a full range of metadata.

Only tools in the first category, those able to extract metadata from a range of materials, are discussed below. Enhancements to the PRONOM service of The National Archives (UK) may, in the future, assist in locating tools capable of extracting metadata from single specific formats.

A3.2 DROID (Digital Record Object Identification)

Available from The National Archives (UK) -
http://www.nationalarchives.gov.uk/aboutapps/pronom/tools.htm

DROID is a platform-independent Java-based application which identifies the format and version of files based upon comparison of file data streams against a set of known signature byte sequences. The signature byte sequences are held in a signature file, which may be updated automatically from The National Archives web site by the DROID application. In March, 2006, Version 9 of the signature file contained signature byte sequences for 57 named file formats (including 159 versions of those formats), and a further 387 tentative file format indicators based on file extension alone.

The main function of DROID is to identify a wide range of file formats as conclusively as possible, including versions. Where a number of possible matches are identified, for example, where multiple versions of a format contain the same signature byte sequences, all matches are listed, along with an indication of the degree of match (e.g. Tentative, Positive). DROID may also notify of suspected mismatches between the format as identified by internal signatures and the filename extension.

DROID, identifies a wider range of formats than the other tools noted (JHOVE and the National Library of New Zealand Metadata Extraction Tool), and, where available, indicates the Persistent Unique Identifier (PUID) that has been assigned to the identified format within The National Archives format registry, PRONOM. However, it does not extract any further metadata from files, nor generic metadata about them (e.g. creation date etc.).

DROID could be used by repositories at least to provide file format identity information to fulfil the PREMIS mandatory elements:

- format
    - formatName and/or a combination of formatRegistryName and formatRegistryKey

Further format specific tools for metadata extraction might then be invoked based on format identifications from the DROID output.

Samples of output:

- Audio
- Databases
- HTML
- Image
- Multimedia
- Text
- Video

A3.3 National Library of New Zealand Metadata Extraction Tool

Available from the National Library of New Zealand - http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction

The National Library of New Zealand Metadata Extraction Tool is also a platform-independent Java-based application, designed to extract preservation metadata from a range of formats. Metadata may be extracted for each format by a specific modular "adapter", and can be output to XML in either an "adapter-native" schema or in a schema complying with the National Library of New Zealand's Preservation Metadata scheme. The tool is designed to be extensible, allowing creation of additional adapter plug-ins by other parties and the structuring of output via XSLT to suit alternative metadata schemes. The tool is capable of recognising and processing a range of formats and versions of formats, but does not currently appear to validate files against their identified format.

The range of formats which can be recognised and for which metadata can be extracted are currently:

- TIFF, JPEG, GIF, BMP
- WAV, MPD
- HTML, PDF
- MS Word 2, MS Word 6
- Word Perfect
- MS Excel, MS PowerPoint
- MS Works
- Open Office

Although the range of formats for which there are adapters is currently small, these cover file formats that may be commonly encountered, and the amount of metadata that is extracted can be quite extensive, particularly in "native" mode. If a format is not recognised, generic file metadata can nonetheless be collected, such as filename, size and date created. The tool can be run via either a Windows interface or from a command line.

Samples of output:

- Audio
- Databases
- HTML
- Image
- Multimedia
- Text
- Video

A3.4 JHOVE (JSTOR/Harvard Object Validation Environment)

Available from Harvard University Library: http://hul.harvard.edu/jhove/

JHOVE is also a platform-independent Java-based application, primarily designed to identify a range of formats and validate files against their purported formats. It can also recognise format sub-types and versions. In characterising files, JHOVE is also capable of extracting technical metadata from the range of formats and producing XML-encoded or plain text output. JHOVE is also modular and extensible in design, allowing creation of additional modules as needed.

There are currently modules available for characterisation of 12 main format types, comprising around 52 versions or distinct subtypes of those formats. The main formats recognised and for which metadata can currently be extracted are:

- TIFF (including DNG), JPEG, JPEG200, GIF
- WAV (including BWF), AIFF
- HTML, XML
- ASCII, UTF-8
- PDF (including PDF/A)
- "Bytestream" (always valid)

If a format is not recognised, it is classed as a "bytestream" and is always well-formed and valid. The tool can be run via either a Windows interface or from a command line.

The metadata extracted can be quite extensive. For images and audio, XML output can be generated according to the MIX schema for still images and the Audio Engineering Society (AES) schemas for audio objects and time code formats.

Again, not all the formats to be accepted by APSR repositories are recognised by JHOVE, and other tools may also be required.

Samples of output:

- Audio
- Databases
- HTML
- Image
- Multimedia
- Text
- Video

A3.5 Summary of functions covered by tools

Tool	Identify format (Tentative)	Identify format (Confirm)	Identify versions	Validate format	Collect generic file MD	Collect material type MD	Collect file format MD
DROID	Yes [546 formats]	Yes [159 formats]	Yes	No	No	No	No
NLNZ-MET	Yes [15 formats]	(Some)	(Some)	No	Yes	Yes	Yes
JHOVE	Yes [52 formats]	Yes [52 formats]	Yes	Yes	Yes	Yes	Yes

- A comparison of formats recognised by the tools with respect to the list of supported and commonly used formats in Appendix 2

A3.6 Recommendations

1. Repositories identify and validate the file formats of objects on ingest or shortly thereafter.
2. Repositories use tools on ingest of an object or periodically on new objects, to collect extra metadata and/or metadata which can't be easily supplied during submission.
3. The National Library of Australia embark on a more detailed audit to align metadata able to be output against recommended preservation metadata elements and make results of this work available when ready.

Appendix 4: Gap reports for ANU DSpace and UQ Fez/Fedora repositories

This analysis was current at 19 May 2006. The reports look at the level of support for the core preservation metadata elements (i.e. PREMIS semantic units) and include recommendations for enhancements where gaps were identified.

A4.1 ANU DSpace repository

PREMIS semantic unit	Supported?	Comments on current level of support	Possible enhancements
Object Identifier	Supported	Items are given globally unique Handles. Files (DSpace bitstreams) are given a local database identifier only.	DSpace are planning to use infoURIs for bitstreams which would be globally unique.
Preservation Level	Supported	DSpace has 3 support levels (Supported, Known, Unsupported) but ANU doesn't assign them. Can be defaulted from the file format.	Content policy development around these levels. The number of levels could be increased if necessary to conform with a generic set of service levels.
Object Category	Supported		
Composition Level	Not supported; not applicable	Default would be 0 for all files in supported formats. Files that have a composition level of higher than 0 (e.g. zip files) would fall into the category of unknown, unsupported format.	
Fixity	Supported	DSpace calculates checksum on ingest.	Checksum checker is coming in next version of DSpace (v1.4)
Size	Supported	DSpace records size on ingest.	
Format	Supported	DSpace determines this from the filename extension. Format version is not determined at present.	Format validation by running a tool such as JHOVE or DROID over the repository or on ingest. Tools could also determine format version.
Significant Properties	Not supported; not applicable		If required, submission forms could be modified to ask for this information.
Inhibitors	Not supported; not applicable	Policy would be not to support files to which this applies.	
Creating Application	Not supported	ANU policy is to avoid providing preservation level support for formats where creating application is important, and instead promote popular, open formats. The date the file was originally created not supported unless explicitly provided in the metadata being submitted.	Present or future tools may be able to provide this information automatically. If required, submission forms could be modified to ask for this information. Descriptive metadata may indicate date of creation for born digital items.
Original Name	Supported		
Storage	Supported		
Environment	Not supported	Not an issue for supported formats.	Global format or environment registries (under development) will meet this need. If required for special cases, submission forms could be modified to ask for this information.

Signature Information	Not applicable		
Relationship	Not supported	Only supported currently through DC.Relation, which is at the item, not the file level.	Relationships including structural maps could be stored as a serialised bitstream with the object.
Linking Event	Partially supported	Ingest event can be determined from database. History logging module exists though it doesn't work properly, has performance issues, and is not being used. The checksum checker will be separate from the history module and the logging systems for the database (eg editing and viewing) are also separate. Theoretically events could be got from logs but it might not be easy.	Fixity check logging will be possible in next version of DSpace. If JHOVE or DROID are run over the repository, the validation event could be determined from the JHOVE or DROID output stored with the object. Ideally the logging systems should be integrated and work properly to record events and their outcome for a particular object.
Linking Intellectual Entity (Descriptive metadata)	Supported	Each item has a qualified Dublin Core record. Other descriptive metadata may be held in serialised bitstreams.	
Linking Permission Statement	Supported	Some rights may be stored in DC.Rights. Licences (including Creative Commons) may be stored with the object.	

A4.2 University of Queensland Fez/Fedora repository

PREMIS semantic unit	Supported?	Comments on current level of support	Possible enhancements
Object Identifier	Supported	Persistent identifier (PID) at item level is UQ prefix followed by a number assigned by Fedora. Datastreams (files) associated with an item are identified by their filenames. infoURIs containing the PID and filenames can be constructed.	
Preservation Level	Not supported	Haven't needed it yet. Could be defaulted to a single level. They have just received a request for quotation for repository services. There is nowhere specific to store service levels.	Could add field to descriptive metadata form or store service level agreement as datastream with the object.
Object Category	Supported	Preservation metadata derived from JHOVE is stored at file level.	
Composition Level	Not applicable	Default would be 0 for all files in supported formats.	
Fixity	Not supported	Checksums are not being generated.	Checksums should be generated on ingest and stored.
Size	Supported	Is in JHOVE metadata.	
Format	Supported	Is in JHOVE metadata. Includes version.	
Significant Properties	Not supported	Could be stored in description.	If required, could be added to the submission forms.
Inhibitors	Not applicable		
Creating Application	Not supported	JHOVE doesn't do application names and versions (but does get camera names for JPEGs). Original creation date of files not stored either.	If required, could use another tool which detects versions. Could be added to submission forms. Descriptive metadata may indicate date of creation for born digital items.

Original Name	Supported	File keeps its original name unless it doesn't conform to NCName. Not sure if original name is kept in this case.	
Storage	Supported		
Environment	Not supported	Not an issue for supported formats.	Global format or environment registries (under development) will meet this need. If required for special cases, submission forms could be modified to ask for this information.
Signature Information	Not applicable		
Relationship	Not supported	Fedora RELS-EXT is being used for relationships to other items. RELS-INT for relationships between datastreams is in the current version of Fedora but Fez is not using it yet. Internal relationships are only implicit through filenaming conventions at present - there is no metadata about relationships.	Implementation of Fedora's RELS-INT.
Linking Event	Not supported yet	Fedora has some audit trail recording. Fez is currently being developed to use this. History logging could be done automatically or manually.	Continued implementation of history logging.
Linking Intellectual Entity (Descriptive metadata)	Supported	Dublin Core record. Additional descriptive metadata may be stored.	If required, additional fields can be added to submission forms.
Linking Permission Statement	Supported	Fez has sophisticated and flexible rights management. Roles and groups (eg Fez groups, Shibboleth groups, targeted IDs) can be linked to different actions.	

A4.3 Recommendations

1. ANU and UQ repositories consider implementing the enhancements suggested in the gap reports, particularly
    - recording of preservation events
    - recording of structural relationships
    - file format validation (ANU)
    - checksum generation (UQ)
2. ANU and UQ repositories particularly take note of the functional requirements in Appendix 6 when implementing the recording of preservation events.

Appendix 5: Submission models for key digital content categories

The aim of this product was to look at workflow models for different types of digital content e.g. electronic publishing, digitisation of physical object, and to recommend how metadata should be acquired and what metadata a SIP should contain.

At this stage it was felt not to be appropriate to develop submission models for ANU and UQ as their systems already have underlying data models and submission processes, both established and under development. This project has specified in Appendices 1, 4 and 6 what metadata is required, but leaves decisions on how to enhance systems to collect it to the repository administrators and developers. The National Library of Australia has been reviewing the architecture of its Digital Collections Manager and may in future develop general submission models which may be useful to other repositories.

Regardless of the type of digital content, the main methods of submission involve:

- one-at-a-time submission using a web form to collect metadata
- batch submission using a web form to collect metadata
- batch submission contains accompanying metadata
- harvesting including accompanying metadata

In each case a SIP is compiled which the repository can ingest. In the first two cases, the SIP is compiled after the web form is completed. In the latter two cases, the batch or harvested submission may already be in the form of a SIP compiled by an external workflow system or tool. Other APSR projects are developing examples of these tools e.g. in the Bidwern project and FIDAS (Fieldwork Data Sustainability) project (the tool is called FieldHelper). Among other things, these tools help researchers organise and tag their files, then automatically prepare the data for uploading to institutional repositories, for instance, by compiling SIP packages as METS documents for ingest to DSpace. Work is also being done with electronic journal publishing systems.

Whatever method is used for the actual submission, the aim should be to capture as much metadata as possible (automatically where possible) as a by-product of creating a digital object.

Appendix 6: Preservation Event use cases and functional requirements

A6.1 Introduction

This document describes the requirements for

- actions that need to be taken on objects in a digital preservation repository
- recording those actions or events.

Lavoie says about actions in the OAIS Functional Model:

> " ..the Archival Storage function is responsible for ensuring that archived content resides in appropriate forms of storage ... and that the bit streams comprising the preserved information remain complete and renderable over the long-term. To meet this responsibility, Archival Storage periodically undertakes procedures such as media refreshment or format migration. The Archival Storage function also implements various safeguard mechanisms, such as error-checking procedures, to evaluate the outcome of preservation processes, as well as disaster recovery policies to mitigate the effects of catastrophic events .."

The PREMIS Data Dictionary says about documenting events:

> "An Event is an action that involves at least one object or agent known to the preservation repository." "Documentation of actions that modify (that is, create a new version of) a digital object is critical to maintaining digital provenance, a key element of authenticity." "Even actions that alter nothing, such as validity and integrity checks on objects, can be important to record for management purposes."

These requirements are primarily concerned with events in the above context, that is, actions, relevant to preservation, on "master" or archival copies of objects. It is recognised that repositories usually have other purposes in addition to preservation and that display copies, supporting files, metadata etc may exist in repositories as digital objects in addition to the archival "content" object. A repository may log actions and events for various purposes. The requirements listed here may therefore only be a subset of an individual repository's requirements.

These requirements are deliberately generalised in order to be applicable to any repository, regardless of any particular software, implementation or architecture. Repository administrators and developers would need to determine more specifically how the requirements would be implemented in their repositories.

The use cases below apply both to actions that are performed on a single object and actions that are performed on a batch of objects (or all objects) in a repository.

A6.2 Use cases

1. **Performing an action on an object which doesn't change the object** e.g. error checking.
2. **Performing an action on an object which transforms an object into a new object (without materially changing its content)** e.g. migration to a newer format.
3. **Deleting an object**
4. **Updating the content of an object:** An action performed in some repositories, not usually for preservation purposes, but included for clarification.

5. **Updating metadata about an object:** It is desirable from a preservation point of view to have the most complete, accurate metadata available, therefore there needs to be a way of updating the metadata as new information comes to light.

A6.3 Actors

These are the Actors (roles) in the use cases below. The Actors are "systems" but these may be manual systems (i.e. people), automated systems or a mixture.

- **Preservation Monitor:** System that monitors preservation policies and risk and provides alerts when preservation action needs to be taken.
- **Workflow System:** System used by repository administrators (human users) to interact with the other systems.
- **Event Manager:** System that performs actions on objects in the Repository.
- **Repository:** System that stores and manages objects and metadata.

A6.4 Use case 1: Performing an action on an object which doesn't change the object

This use case applies to, for instance, PREMIS eventType

- digital signature validation
- message digest calculation (checksum calculation)
- fixity check (checksum check)
- validation (of format e.g. a file complies with the format specification its filename implies)
- virus check

Message digest calculation and format validation, and if applicable, fixity check and virus check, should ideally be done on ingest of an object. In this case they may or may not be recorded as separate events but if not, they should be noted in the event details or in repository policy and procedures.

If not done at ingest, these events may occur some time later, when they may be recorded as separate events.

Trigger: Identification of preservation risk (by a person or preservation monitoring system) or part of an auditing process (one off or regularly scheduled)

1. The Preservation Monitor or Workflow System alerts the Event Manager that an action needs to be performed on an object.
2. The Event Manager schedules the event.
3. The Event Manager peforms the action.
4. The Event Manager notifies the Repository that an event has taken place along with details of the event.
5. The Repository records the event details. (see Event Details below).

A6.5 Use case 2: Performing an action on an object which transforms an object into a new object without materially altering the content.

This use case would apply to, for instance, PREMIS eventType

- migration (to another format)
- normalization (to a standard or supported format)
- compression

An event which changes the preservation copy of an object should always be recorded.

Trigger: Identification of preservation risk (by a person or preservation monitoring system) or implementation of a policy decision e.g. to migrate all files of a certain format to a newer, better supported format.

Base course: A new object is created and the old object is kept.

1. The Preservation Monitor or Workflow System alerts the Event Manager that an action to change an object needs to be performed.
2. The Event Manager schedules the event.
3. The Event Manager takes a copy of the object, and modifies it to create a new object.
4. The Event Manager submits the new object to the Repository along with details of the event which created it, including its relationship to the old object.
5. The Repository ingests the new object, records the relationship between the new and old objects, applies version information (especially if the new object is the new master archival copy) and assigns a unique identifier to the new object.
6. The Repository stores relevant preservation metadata about the new object.
7. The Repository ensures descriptive, rights and and any other relevant metadata from the old object are associated with the new object.
8. The Repository records details of the event which created the new object and associates the event with the new and old objects.
9. The Repository records the event which ingested the new object. If the ingest event is not stored explicitly the details must be able to be output to conform with the draft APSR METS profile (Appendix 8).

Alternative course: A new object is created and the old object is not kept.

After step 9:

- The Repository removes the old object from the repository but keeps the object identifier and preservation metadata needed to trace an object's provenance: at least the object's format.
- The Repository records the deletion of the old object.

A6.6 Use case 3: Deleting an object from a repository

Repositories will have their own policies on the circumstances where deleting an object is allowed. It is expected that some metadata about an object will be kept even though the object itself is removed. This should be at least the object identifier and some descriptive metadata (or a link to it e.g. through a relationship with a current object).

Trigger: Implementation of a policy decision to delete an object. For instance, a decision to delete all objects of a certain type e.g. non-current versions of masters, or a policy to only keep certain objects for 10 years.

1. Workflow System (after checking its rules about which objects can be deleted and by whom) or Preservation Monitor instructs the Repository to delete an object.
2. The Repository checks whether the object to be deleted is part of the provenance history of the current "master" copy of an object. (Depending on the particular Repository, this may have been recorded in a provenance history, or through a relationship (direct or indirect) with the master object, or through an event or chain of events that led to the creation of the current master object.)
3. If it is, the Repository should have rules about what preservation metadata needs to be kept (this should include at least the object's format). The repository administrator should be able to configure these rules.
4. The Repository checks any other relationships, links or associations the object has. The Repository will have rules to deal with these before or when an object is deleted in order to maintain the integrity of the data.

5. The Repository deletes the object and keeps any metadata required from the above checks.
6. The Repository records the deletion event.

A6.7 Use case 4: Updating the content of an object

An example of this use case is a depositor changing the content of a document.

Particularly for internal documents, reports etc, the Workflow System may well assign a version number to the new document. However although this can be regarded as a new "version" of the old object, it is different from Use case 2 above. The repository should differentiate between different versions of the same content, and "versions" where the content is not the same.

Instead, this new "version" should be regarded as a new "work" (PREMIS Intellectual Entity) with a relationship to the old "work". It should have its own descriptive metadata distinct from the descriptive metadata of the old work, similar to the way different editions of a book have their own records in a library catalogue.

This new work should be able to have its own preservation policy. For example, the latest "version" may need to be kept indefinitely, whereas the earlier "versions" may only need to be kept for a defined period. Or the policy may be to only keep the latest "version" (which has been authorised through the Workflow System before submission to the Repository) and delete previous "versions" immediately.

Trigger: The depositor may use a Workflow System to take a copy of the object in the repository, edit it and re-submit it, or the depositor may edit their own local copy of the original and submit it through the Workflow System as a new "version" of the original object.

1. Workflow System accepts the new object.
2. Workflow System submits the new object to the Repository. The Workflow System may provide a complete OAIS SIP, or only information that is different for this object.
3. The Repository ingests the new object, assigns a unique identifier to the new object and records the relationship between the new and old objects. This relationship may be recorded through the descriptive metadata only or may be more explicit in the system.
4. The Repository may associate updated descriptive, rights and and any other relevant metadata from the old object with the new object, if the Workflow System only provides updated information.
5. The Repository stores relevant preservation metadata about the new object.
6. The Repository records the event which ingested the new object. If the ingest event is not stored explicitly the details must be able to be output to conform with the draft APSR METS profile.

A6.8 Use case 5: Updating metadata about an object

For example, the descriptive metadata about a photograph may need to be changed when new information comes to light about the people depicted in it.

This use case may be applied to administrative, structural etc as well as descriptive metadata.

It is up to individual repositories whether only one version or different versions of the metadata are kept, or even if a record of changes is kept. From a preservation point of view the authenticity of the content object is most important and keeping a record of changes to the content object is mandatory, but it is optional for the metadata. Whether or not to keep previous versions of the metadata depends on its significance and what it might be used for. It is however usual to keep at least the date the metadata was originally created and by whom (organisation rather than person) and the date it was last updated and by whom.

1. Workflow System accepts new version of the metadata.

2. Workflow System sends new version of metadata with the object identifier to the Repository.
3. The Repository stores the new version of the metadata as the current version and associates it with the object.
4. The Repository may or may not keep previous versions of the metadata.
5. The Repository keeps the date the first version of the metadata was created and updates the date last updated to today's date. The Repository may or may not keep other dates.

A6.9 Event Details

What details are recorded about an event and how they are stored will depend on the particular Repository's data model and architecture.

For the purposes of publishing or exchanging metadata about an archival object, the Repository should be able to conform to the proposed APSR METS profile. This profile specifies that a history of events describing an object's provenance be output in digiprovMD using the schema for the PREMIS Event Entity (see the PREMIS Data Dictionary.)

These are the semantic units of the PREMIS Event Entity (NR=not repeatable; R=repeatable; M=mandatory; O=optional):

- eventIdentifier NR M
  - eventIdentifierType NR M
  - eventIdentifierValue NR M
- eventType NR M
- eventDateTime NR M
- eventDetail NR O
- eventOutcomeInformation R O
  - eventOutcome NR O
  - eventOutcomeDetail NR O
- linkingAgentIdentifier R O
  - linking AgentIdentifierType NR M
  - linkingAgentIdentifierValue NR M
  - linkingAgentRole R O
- linkingObjectIdentifier R O
  - linkingObjectIdentifierType NR M
  - linkingObjectIdentifierValue NR M

The profile also says that additional information about agents associated with events may optionally be recorded. Agents may be persons, organisations or software.
The PREMIS Agent Entity has the following semantic units:

- agentIdentifier R M
  - agentIdentifierType NR M
  - agentIdentifierValue NR M
- agentName R O
- agentType NR O

Even if there is no additional information for software or a device, placing it in Agent in a document to conform with the draft APSR METS profile will facilitate mapping in the receiving repository's database.

Additional information considered useful but not covered by PREMIS should be recorded. Other more detailed schemas to describe events may emerge and/or PREMIS Event may be enhanced in the future

A6.10 Recommendations

1. ANU and UQ repositories particularly take note of the functional requirements for preservation events in Appendix 6 and bring them to the attention of their open source communities as they begin to develop event logging functionality.

Appendix 7: Issues / enhancements to PREMIS and existing schemas and protocols that might be used.

A7.1 PREMIS conformance

This project is using PREMIS in two ways:

- firstly as a checklist to identify gaps in and make recommendations about metadata which should be collected;
- secondly as a container within a METS profile for metadata exchange in the scenario of transferring custody of an object from one repository to another. The draft profile maps PREMIS Object to techMD and PREMIS Event (and Agents if necessary) to digiprovMD.

The APSR repositories will aim to be PREMIS conformant in as far as being able to produce a METS document with metadata in a container using a PREMIS namespace valid according to the PREMIS xml schemas. If the data were not available they would have to be included with values of "unknown" or "not applicable". However in this case, i.e. if the repository could not supply a real value for a mandatory semantic unit, they could be regarded as not PREMIS conformant.

A7.2 Issues encountered in PREMIS

Some issues encountered while examining the PREMIS Data Dictionary were raised with the PREMIS Implementors' Group and added to the errata for fixing in the next version of the data dictionary e.g.

- For Representations objectCharacteristics is Not Applicable, but significantProperties, within the container ObjectCharacteristics, is Applicable. significantProperties may have to be moved out from objectCharacteristics.
- swVersion is Not Repeatable for Representation and Bitstream but is Repeatable for File. It should be Not Repeatable for File as well.
- relatedObjectSequence is mandatory even though it doesn't apply to derivative relationships (not mandatory in schema but is in Data Dictionary)
- relatedEventIdentification:This obligation means that a relationship must have an event, but this only applies to derivation relationships, not structural relationships. This had already been noted as an error on the PIG list.

An interpretation issue was found with "relationship" in the PREMIS Data Dictionary :

STRUCTURAL RELATIONSHIPS: Under relationshipType on page 2-62 it says "structural=a relationship between parts of an object". This accords with what PREMIS says on page 1-8 i.e. structural relationships are about how to put back together a digital object which consists of more than one part or file. However the paragraph under Derivation relationships on page 1-9 says "A structural relationship among objects can be established by an act of derivation before the objects were ingested by the repository ... " and "..They do not have derivation relationships with each other, but do have a structural relationship as siblings (children of a common parent)". It's confusing to describe this as a structural relationship because the 'siblings' are not part of the same digital object - they belong to different representations.

"PARENT" AND "CHILD": On page 2-63 it says "is child of = the object is directly subordinate in a hierarchy to the related object ..." and "is parent of = the object is directly superior in a hierarchy to the related object ...", but it doesn't say what the hierarchy relates to. In the paragraph (on page 1-9) referred to above, "parent" refers to the object from which the "children" are derived, whereas on page 6-5 "children" is used to describe components of a web site. In the former case the parent has a "source of" relationship with the children; in the latter case the children have an "is part of" relationship with the (parent) website. In NLA's Digital Collections Manager system the term

"child" is used to denote "part of" at the Intellectual Entity level. Because "parent" and "child" can be used in various contexts, it is recommended to avoid "is parent of", "is child of", "has child" and "has parent" in relationshipSubType and that more precise terms such as "source of", "derived from", "is part of", "has part" be used instead.

Allowing reciprocal relationships to be described in two places can give rise to data integrity problems e.g. one object may have an "is part of" relationship to a second object, which may in turn have an "is part of" relationship to the first object.

Integrity problems could also arise because two way linking is allowed between Object and Event. An Object can be linked to an Event through the Object's semantic units "relatedEventIdentification" or "linkingEventIdentifier"and an Event can be linked to an Object through the Event's semantic unit "linkingObjectIdentifier".

Another issue with Events and linkingObjectIdentifier is that there is no way of saying (if applicable) which was the "source" object and which the "output" object other than by referring back to one of the objects to find its relationship to the other object, and again there is the potential for inconsistency.

A7.3 Proposals for enhancements to PREMIS

At this stage, other than fixing issues in the first version of the data dictionary, no particular enhancements have been identified. However once repositories begin to use PREMIS desired enhancements will probably be identified.

Proposals for enhancements will be sent for discussion to the PREMIS Implementors' Group list and will be formally submitted to the Editorial Committee for the PREMIS Maintenance Activity, which the National Library of Australia has been invited to join.

A7.4 Other existing schemas and protocols

Other schemas and protocols recommended in this report are

- METS as a format for *metadata exchange* (SIP, DIP) and possibly as a storage format for (AIP) as well. METS was chosen because it is the best understood schema and is the one the PREMIS Implementors' Group have discussed.
- MODS as a common exchange format for *descriptive metadata* because it is more granular than Dublin Core but is still a general-purpose and easy to understand schema
- Extension schemas recommended by METS for *file-format specific metadata* (see draft METS profile in Appendix 8 for more detail). Appendix 1 (Preservation metadata elements) and Appendix 8 (draft METS profile) recommends some "extensions" to these extensions.
- *Access rights metadata* was not studied in detail. Some possibilities are PREMIS Rights, METS Rights, Creative Commons licences and XACML.

Again, once repositories begin to use these, issues and enhancements will be identified and conveyed to the groups responsible.

An issue with using multiple schemas is they overlap, leading to redundant metadata and an element may be mandatory in more than one schema. For example there is an element for "checksum" or equivalent in METS, PREMIS and MIX. As JHOVE outputs MIX metadata for images and AES metadata for audio, crosswalks will need to be developed to map elements from these schemas to PREMIS where an equivalent element exists in PREMIS.

A7.5 Recommendations

1. Australian repositories, particularly the National Library of Australia, continue to actively participate in development of standards relevant to digital preservation.

2. Australian repositories continue to actively participate in development of open source software for digital repositories, encouraging support for digital preservation metadata and standards in these developments.
3. The National Library of Australia develop crosswalks, if not already available, to map elements from schemas output by automated tools to PREMIS where an equivalent element exists.

Appendix 8: Proposed profile for exchanging metadata

A8.1 Introduction

This Appendix proposes a METS profile for exchanging metadata about digital objects. The proposed profile is in the form of a table of rules and recommendations. It will be revised after testing with ANU and UQ and further consultation. It can then be expressed in XML using the formal METS profile schema and submitted to METS for registration. The maintainer of the profile will be the National Library of Australia.

The purpose this profile addresses is use of a METS document to transfer custody of a digital object or set of digital objects from one repository to another. This is because this scenario requires the most complete set of preservation metadata. In the OAIS model the transferring repository produces the METS document as a DIP which becomes a SIP in the receiving repository. In some repositories such a document may also be stored and constitute an AIP.

SIPs for material being submitted to a repository for the first time and DIPs produced for purposes other than transferring custody of an object, may be based on this profile, containing a subset of the metadata specified here.

It is a generic profile meant for use among Australian repositories, particularly members of the Australian Partnership for Sustainable Repositories. It is not specific to a particular system or implementation. Repositories will need to map their implementation specific requirements or profile to this common one.

A8.2 General notes

A conforming METS document represents a discrete item of interest for access and preservation purposes. An item has a discrete set of metadata to describe its content.

An item will be completely described in a conforming METS document, therefore in the case of an item with parts, the parts will be completely described within the document as well.

An item (whether it contains parts or not) may be part of another item. A conforming METS document may therefore represent an "item" or a "part", as long as the "part" is a discrete item of interest with its own descriptive metadata.

A conforming METS document would not describe a "collection" or large sets of items. Nor would conforming METS documents be expected to contain information about an item's relationship to a collection or to other items in the collection, other than through the descriptive metadata. However a non-conforming METS document for the collection could contain pointers to conforming METS documents for items in the collection.

A conforming METS document must contain the files or pointers to the files comprising the archival copy of an item, as well as all supporting files and metadata necessary for its long term preservation and access. A conforming METS document may contain files or pointers to files comprising other representations of the item (e.g. thumbnail copy, display copy) as well, along with sufficient metadata to render or execute the files properly.

All available metadata should be included for the archival copy of an object - there should be no loss of granularity. If a namespace or schema doesn't cover all the metadata, the extra metadata should be included under a local namespace. If a receiving repository cannot process some of the metadata (whether the metadata is specified in this profile or not), the receiving repository should store the metadata in its raw xml form or store the whole METS document, rather than discard any metadata. The repository may be able to use the metadata eventually (e.g. if the system is

enhanced) and if not, a human could read it and hopefully make some sense of it, if it became necessary for problem solving or to answer a query about an item.

The order of precedence followed in this profile for placing metadata is METS, PREMIS, other schemas specified in this profile, any other schemas. That is, metadata should be placed in a METS element or attribute if possible; if there is no appropriate place in METS, it should be placed in a PREMIS element if possible; otherwise use an element from a recommended schema; otherwise use another established schema if possible.

A8.3 Schemas

METS extension schemas:

MIX
http://www.loc.gov/standards/mix/mix.xsd
MODS
http://www.loc.gov/standards/mods/v3/mods-3-2.xsd
PREMIS
http://www.loc.gov/standards/premis/v1/PREMIS-v1-1.xsd
textMD
http://dlib.nyu.edu/METS/textmd.xsd

Other schemas:

AMD: LC-AV Audio Metadata Extension Schema
http://lcweb2.loc.gov/mets/Schemas/AMD.xsd
VIDEOMD: LC-AV Video Metadata Extension Schema
http://lcweb2.loc.gov/mets/Schemas/VMD.xsd

A8.4 Tables of METS elements and attributes

This table is based on that used by MacKenzie Smith in the draft DSpace METS profile posted to the PREMIS Implementors' Group.

R=Repeatable NR= Not Repeatable M=Mandatory O=Optional

A8.4.1 <mets> element group

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
<mets>	NR M	Must contain PROFILE attribute and a <metsHdr> element.
PROFILE	NR M	Optional in METS. The value for this attribute will be: National Library of Australia METS SIP Profile 1.0
OBJID	NR M	Optional in METS. Must have a primary identifier assigned to the METS document. It should be unique within the repository but does not have to be globally unique.
ID, LABEL, TYPE	NR O	No recommendations.

A8.4.2 <metsHdr> element group

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
<metsHdr>	NR M	Must contain CREATEDATE and LASTMODDATE attributes.
CREATEDATE	NR M	Optional in METS but mandatory in this profile.

LASTMODDATE	NR M	Optional in METS but mandatory in this profile.
ID, RECORDSTATUS	NR O	No recommendations.
-<agent>	R M	There must be one instance for the organisation that produced the METS document and one instance for the software and version that produced the METS document. Other agents are optional.
ROLE	NR M	Required in METS. Use "CUSTODIAN" for the organisation and "EDITOR" for the software and version. "CREATOR" may be used for the person responsible, if any.
--<name>	NR M	Must contain the name of the organisation or software and version as appropriate.
--<note>	R O	No recommendations.
-<altrecordID>	R O	No recommendations.

A8.4.3 <dmdsec> element group

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
<dmdSec>	R M	The dmdSec is reserved for bibliographic description and subject analysis of the item and its constituent files, at a ratio of one dmdSec for each unique metadata record.
		Multiple metadata records describing the same item or part using different schemas should be captured in separate dmdSecs and linked via the GROUPID attribute.
		At least one dmdSec with the metadata record for the entire item must be present, the metadata in this dmdSec must conform to the MODS XML schema (one of the METS endorsed extension schemas): mailto:http://www.loc.gov/standards/mods/v3/mods-3-2.xsd
		It is strongly recommended to include additional more granular metadata records (using other schemas or namespaces) if available.
		Each dmdSec must contain an <mdWrap>.
ID	NR M	Required by METS
GROUPID	NR O	Use to identify multiple metadata records (using different schemas) describing the same item or part.
ADMID, CREATED, STATUS	NR O	No recommendations.
-<mdWrap>	NR M	See mdWrap element group section.
-<mdRef>	-	Not supported in this profile.

A8.4.4 <mdWrap> element group within dmdSec or amdSec elements

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
<mdWrap>	NR M	
MDTYPE	NR M	METS requires the presence of this attribute and restricts the values to: MARC, MODS, EAD, DC, NISOIMG, LC-AV, VRA, TEIHDR, DDI, FGDC, LOM, PREMIS, OTHER. Support for MODS in dmdSec and PREMIS in AMDSec are required in this profile, though it is recommended to support others in the above list if applicable to the types of material in the repository.
OTHERMDTYPE	NR O	Use if and only if MDTYPE value is "OTHER".
<xmlData>	NR M	A schema or a namespace is required by this profile. An established schema or namespace is preferred; if not available, a local namespace can be used.

A8.4.5 <amdSec> element group

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
<amdSec>	R M	There must be at least one amdSec. Ideally there should be one amdSec for each content file contained or referenced in the <fileSec> element of the METS document but this may not be practical for some situations. There must be only one <amdSec> per file. (An amdSec may contain repeated <techMD>, <sourceMD>, <digiprovMD> and <rightsMD>).
		Each amdSec must contain an ID attribute and at least one <techMD> element.
ID	NR M	The ID attribute is optional in METS but is mandatory in this profile.
-<techMD>	R M	Each technical metadata record using a schema or namespace (eg PREMIS, MIX) should be organised in its own techMD.
		There must be at least one techMD containing a metadata record in <mdWrap><xmlData> conforming to the PREMIS Object schema. The following elements are mandatory in the PREMIS Data Dictionary for objectCategory "file" and are therefore mandatory in this profile for amdSec pertaining to files. (They are not necessarily mandatory in the PREMIS xml schema since they may not apply to all types of objectCategory.) • objectIdentifierType • objectIdentifierValue • preservationLevel • objectCategory • compositionLevel • storageMedium The following elements from PREMIS Object entity are also mandatory in this profile: • formatName • formatVersion • originalName Values of 'not applicable' and 'unknown' are permitted in mandatory elements where data cannot be supplied. It is strongly recommended to include any optional elements in the PREMIS Object schema for which data is available. Note: The following may be included in PREMIS metadata but are not mandatory in this profile because SIZE, CHECKSUM AND CHECKSUMTYPE are mandatory attributes for the METS <file> element in this profile. • messageDigestAlgorithm • messageDigest • size
		preservationLevel: Use one of the following values: "supported" - fully supported "known" - not supported yet but high priority to try and fully support.

		"unsupported" - known or unknown format, preserve bitstream as is but low priority for support "not_applicable" - not a preservation copy of the item
		For still image files, additional metadata not covered by PREMIS should be encoded using the MIX schema (one of the METS endorsed extension schemas): http://www.loc.gov/standards/mix/mix.xsd
		Additional metadata for text files not covered by PREMIS should be encoded using the schema at http://dlib.nyu.edu/METS/textmd.xsd (one of the METS endorsed extension schemas) with extensions recommended by the National Library of Australia at http://www.nla.gov.au/?
		Schemas for other types of files have not been endorsed by METS yet. Until then, additional metadata for audio and video files should be encoded using schemas proposed for use in the Library of Congress Audio-Visual Prototyping Project. Audio schema is at http://lcweb2.loc.gov/mets/Schemas/AMD.xsd and the video schema is at http://lcweb2.loc.gov/mets/Schemas/VMD.xsd. Additional elements recommended are appended to this document and use the namespace at: http://www.nla.gov.au/?(This namespace is not accompanied by a DTD or schema.)
-ID	NR M	Required by METS
--<mdWrap>	NR M	See mdWrap element group section.
-<rightsMD>	R O	<rightsMD> is optional but where present, it must contain one <mdWrap><xmlData> element, which may contain one of : • METS rights record (using the METS endorsed schema at http://cosimo.stanford.edu/sdr/metsrights.xsd) • Creative Commons distribution licences • PREMIS Rights record (schema at http://www.loc.gov/standards/premis/v1/Rights-v1-1.xsd) • XACML
-ID	NR M	Required by METS
--<mdWrap>	R M	See mdWrap element group section.
-<sourceMD>	R O	May be used but is not required if the dmdSec describes the original source material used to create the METS object e.g. if the METS object is a digital surrogate for a physical item. May be used to describe source materials between the original and current object where the source materials are not digital objects. This profile makes no recommendations about the form this metadata should take. Must be used if <digiprovMD> includes PREMIS event metadata which has a linkingObjectIdentifier to an object which is not being transferred as part of this METS document. In this case <sourceMD> must contain a <premisObject:object>. For example, if a PDF was created from a Word document and the PDF is being transferred but the Word document is not (the Word document may have already been discarded by the transferring repository), the Word document would be described in <sourceMD> as a PREMIS Object.
-ID	NR M	Required by METS
--<mdWrap>	NR M	See mdWrap element group section.
-<digiprovMD>	R M	There must be at least one <digiprovMD> for the current archival or master copy, describing the ingest event into the transferring repository. <digiprov> is optional for objects which are not the master copy. There should be only one <digiprovMD> for each object for which events are recorded. Each <digiprovMD> should only have one <mdWrap MDTYPE="PREMIS"> which has only one <xmlData> element containing all PREMIS Event and

		Agent metadata for the object. Additional <mdWrap><xmlData> elements describing the same events in non-PREMIS schemas may be included but receiving repositories may not be able to process them. Each event must be contained in a separate <premisEvent:event> element with xml data conforming to the PREMIS Event schema. (http://www.loc.gov/standards/premis/Event-v1-0.xsd) Each agent (where agent is recorded) must be contained a separate <premisAgent:agent> element in the same <xmlData> element as the <premisEvent:event> with which it is associated.
		As complete a provenance history as possible should be provided for the 'master' or archival object, describing events (in separate <premisEvent:event> elements) which led to the creation of the current object and its ingest in the transferring repository. This includes changes to the object originally deposited (note that in PREMIS, an object cannot be modified: an event which modifies an object creates a new object) and changes of custody.
		Other types of events occurring after ingest of the current object into the transferring repository may be recorded in additional <premisEvent:event> elements (e.g. format validation, checksum checking)
		The following elements are mandatory within a PREMIS Event in this profile: • eventIdentifierType • eventIdentifierValue • eventType • eventDateTime If the event is one which changes an Object, it is strongly recommended to include information about the hardware / software used. Use the following Event elements under linkingAgentIdenfier: • linkingAgentIdentifierType • linkingAgentIdentifierValue • linkingAgentRole The value in linkingAgentIdenfierType can be the name of an external registry or the repository's own name. The value in linkingAgentIdentifierValue should be a unique identifier within the registry or transferring repository if the agent has one, or else simply a unique identifier to this agent within the METS document. The value in linkingAgentRole should describe the agent's role e.g. "scanner". A controlled vocabulary has not been developed for this element yet. If an organisation *other* than the transferring repository was responsible for an Event, that organisation should also be noted in linkingAgentIdentifier.
		If there is a linkingAgentIdentifier, a <premisAgent:agent> element must be present within the same <xmlData> which contains the<premisEvent:event> element with which it is associated through the event's linkingAgentIdentifier. The following elements are mandatory in this profile: • agentIdentifierType • agentIdentifierValue • agentName (optional in PREMIS)

		• agentType (optional in PREMIS. A controlled vocabulary has not been developed for this element yet.)
		If the object is related to another digital object through an Event and the related object is being transferred as well, the Event should contain a premisEvent:linkingObjectIdentifier which matches the related object's premisObject:objectIdentifier in the related object's <amdSec><techMD> element. If the object is related to another digital object through an Event and the related object is not being transferred, the Event should contain a premisEvent:linkingObjectIdentifier which matches a premisObject:objectIdentifier in a premisObject metadata record in <sourceMD>. For example, a Word document may have been transformed into an RTF then to PDF. If only the PDF is being transferred, each event should be described in a separate <premisEvent:event> with linkingObjectIdentifier matching the objectIdentifier in a <premisObject:object> under <sourceMD>. The Word and RTF files would each be described (even if they no longer exist) in a <premisObject:object> element under separate <sourceMD> elements.
-ID	NR M	Required by METS
--<mdWrap>	NR M	See mdWrap element group section

A8.4.6 <fileSec> element group

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
<fileSec>	NR M	All files must be referenced via the fileSec. <fileSec> must contain one or more <fileGrp>.
<fileGrp>	R M	Use this element to bundle files according to the following categories described in the USE attribute: original: The object originally submitted to a repository by the depositor, if it is being transferred and is different from the master. master: The current archival copy (i.e. the one that has the highest priority for long term preservation). It may be the original or a modified version of the original - this should be able to be determined from digiprovMD. There should be one and only one master. access_representation: The copy preferred for public access, if different from the Master. Other Representation: Any other group of content files which can be used to render an object, which are not the original or the master. structural_map: Strongly recommended that this be a single XML file e.g. a SMIL document for multimedia objects, EAD document for manuscripts, or a description of the file directory structure of a complex object. Filenames referred to should correspond to filenames in <file> OWNERID attribute. This filegroup is not necessary if it provides no more information than the <structMap> element. metadata: Extra metadata files can be included. Not necessary if the metadata is completely covered by other sections of the METS document eg dmdSec, structMap. Files in this group should be xml files. licence: Files that contain licences or rights agreements pertaining to the

		object. This filegroup is not necessary if it provides no more information than <rightsMD> element. support: Any other supporting files (needs an example). other: For files which don't fit into the other categories.
		In this profile: <fileGrp> must contain one or more <file> <fileGrp> may not contain any nested <fileGrp> elements.
ID, VERSDATE, ADMID	NR O	No recommendations
USE	NR M	Each <fileGrp> must have a USE attribute with a value from the above vocabulary.
		There must be one and one only <fileGrp> with USE="master"
		There can be 0 or 1 <fileGrp> with USE="original"
		There may be any number, or none in the other categories.
--<file>		See <file> element group section.

A8.4.7 <file> element group

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
<file>	R M	<file> must contain either a single <FLocat> or an <FContent>. This profile doesn't provide for <stream>, <transformFile> or nested <file> at present.
ID	NR M	Required in METS
MIMETYPE	NR M	Optional in METS but required in this profile.
SEQ	NR O	No recommendations.
SIZE	NR M	Optional in METS but required in this profile.
CREATED	NR O	Strongly recommended. Should be the date the creating application created the file, not the date it was ingested in the transferring repository.
CHECKSUM	NR M	Optional in METS but required in this profile.
CHECKSUMTYPE	NR M	Optional in METS but required in this profile. METS specifies the following values: HAVAL MD5 SHA-1 SHA-256 SHA-512 TIGER WHIRLPOOL
OWNERID	NR O	May be used to provide a unique identifier (including a URI) assigned to the file which may differ from the URI used to retrieve the file. Strongly recommended for filenames referred to in files in <fileGrp> with USE="structural_map", or which may be used by a 'root' file to reconstruct or render an object. Must be used to provide a link to the file's administrative metadata.
ADMID	NR M	Must be used to provide a link to the file's administrative metadata.
DMDID, GROUPID, USE	NR O	No recommendations.
-<FLocat>	NR O	Repeatable in METS but not in this profile. An <FLocat> must be provided for each <file> if the content of the file is not embedded in <FContent>. Flocat can only be used *if and only if* the URL can be *guaranteed* under normal conditions (i.e. excluding network/connectivity issues) to

		be accessible by an ingesting party. Any URL either needs to point directly to a METS package or a service which exposes the METS package to the requesting party (receiving repository).
ID, USE	NR O	No recommendations
LOCTYPE	NR M	Required in METS and values restricted to: URN URL PURL HANDLE DOI OTHER "OTHER" must not be used in this profile.
OTHERLOCTYPE	-	Not supported by this profile.
xlink	NR M	No recommendations
-<FContent>	NR O	Must be present if there is no <FLocat>. As specified in METS, the content file must be either Base 64 encoded and contained within the subsidiary binData wrapper element , or consist of XML information and be contained within the subsidiary xmlData wrapper element.
ID, USE	NR O	No recommendations
binData	NR O	No recommendations
xmlData	NR O	No recommendations

A8.4.8 <structMap> element group (see example METS document)

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
<structMap>	NR M	Repeatable in METS but not in this profile. The structMap element must describe the structure of the whole object represented by the METS document.
ID	NR O	No recommendations
TYPE	NR O	No recommendations
LABEL	NR O	No recommendations
-	R M	For an object whose structure is hierarchical the structure should be encoded as a tree of nested elements. For an object consisting of a single file there will be a single element. For a complex (non-hierarchical) object consisting of multiple files, there will be single with the first child <fptr> indicating the 'root' file or the file which 'knows' how to 'get' the other files in order to render the object. If there is no such file, the first <fptr> may point to a file in the <fileGrp> with USE="structural_map". The first level elements must represent the whole object. Lower level elements represent parts of the object (see the example METS document).

A8.4.9 element group

Element / Attribute	Profile occurrence / obligation	Profile rules and recommendations
ID	NR O	No recommendations
ORDER	NR M	The first level elements represent the whole item and must have an order of "1". Lower level elements must have an order attribute, at least one of which must have an order of "1". There may be more than one

		at the same level with the same order number.
ORDERLABEL	NR O	No recommendations
LABEL	NR O	Strongly recommended
DMDID	NR O	Must be present for the first level representing the whole item. Should be present for the lower level s if there is a corresponding dmdSec for that part of the item.
ADMID	NR O	Should be present if there is a corresponding amdSec for the whole or parts of the item that is not file-specific, e.g. rights metadata.
TYPE, CONTENTIDS	NR O	No recommendations
-<mptr>	-	Not supported in this profile
-<fptr>	R M	Must be at least one <fptr> in each . Each <fptr> must contain a FILEID to point to a file in the METS document. The child elements <par>, <seq> and <area> are not supported in this profile.
ID	NR O	No recommendations
FILEID	NR M	Optional in METS but required in this profile.
CONTENTIDS	NR O	No recommendations

A8.4.10 <structLink> element group

Not supported in this profile.

A8.4.11 <behaviorSec> element group

Not supported in this profile.

A8.5 Sample METS document

See samplemetsdoc.xml

A8.6 Recommendations

1. The National Library of Australia continue to develop and test the proposed METS profile for metadata exchange with input from the Australian National University and the University of Queensland and consultation with the wider digital preservation community with a view to registering the profile formally.

Appendix 9: Glossary

AAF
Advanced Authoring Format. A wrapper format for video to hold additional metadata.

AES
Audio Engineering Society.

AIFF
Audio Interchange File Format. An audio file format standard.

AIP
Archival Information Package. A concept from the OAIS Reference Model.

ANU
The Australian National University. http://www.anu.edu.au/

APSR
Australian Partnership for Sustainable Repositories http://www.apsr.edu.au/. The APSR Project aims to establish a centre of excellence for the management of of scholarly assets in digital format.

AVI
Audio Video Interleave. A video file format.

ARC
An archive file format, used by the Internet Archive.

BMP
Windows OS/2 Bitmap Graphics. An image file format with lossy compression.

BWF
Broadcast Wave Format. An extension of the WAV audio format which also specifies the format of metadata.

CSV
Comma-Separated Variables. A format for storing database content.

CODEC
A 'Compressor-Decompressor', 'Coder-Decoder', or 'Compression/Decompression algorithm'.

Creative Commons
A nonprofit organization that offers flexible copyright licenses for creative works. http://creativecommons.org/

DC
Dublin Core. A standard metadata element set used to describe digital materials. http://dublincore.org/

DCM
Digital Collections Manager, part of the National Library of Australia's digital collections infrastructure. http://www.nla.gov.au/digicoll/infrastructure.html

DIP
Dissemination Information Package. A concept from the OAIS Reference Model.

DNG
Digital Negative. Public, archival format for digital camera raw data

DROID
Digital Record Object Identification.

DSpace
Open source digital repository system that captures, stores, indexes, preserves, and redistributes an organization's research data. http://www.dspace.org/

DV-DIF
Digital Video Digital Interface Format.

EBU
European Broadcasting Union http://www.ebu.ch/

EPS
Encapsulated PostScript.

Fedora
Open source software which gives organisations a flexible service-oriented architecture for managing and delivering their digital content. http://www.fedora.info/

FITS
Flexible Image Transport System http://fits.gsfc.nasa.gov/. A format adopted by the astronomical community for data interchange and archival storage.

GIF

Graphic Interchange Format. A image file format with lossy compression format.

ISO

International Organization for Standardization http://www.iso.org/

JHOVE

JSTOR/Harvard Object Validation Environment.

MARC

The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form. http://www.loc.gov/marc/ The name MARC originated as an acronym for MAchine Readable Cataloging.

METS

Metadata Encoding and Transmission Standard, a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library. http://www.loc.gov/standards/mets/

MIX

NISO Metadata for Images in XML Schema. A standard for technical metadata for still images.http://www.loc.gov/standards/mix/

MODS

A schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications.http://www.loc.gov/standards/mods/

MOV

QuickTime movie format.

MP3

MPEG-1 Audio Layer 3, more commonly referred to as MP3, is a popular digital audio encoding and lossy compression format. (MPEG is the acronym for Moving Picture Experts Group).

MXF

Material Exchange Format. A wrapper format for video to hold additional metadata.

NISO

National Information Standards Organization http://www.niso.org/. Anon-profit association accredited by the American National Standards Institute (ANSI).

OAIS Reference Model

Part of an ISO effort to develop archiving standards. The Reference Model for an Open Archival Information System is at http://public.ccsds.org/publications/archive/650x0b1.pdf

ODF

Open Document Format. An open source format for office applications.

OGG

Ogg Vorbis Codec Compressed WAV File. An open source audio file format with lossy compression.

PANDAS

PANDORA Digital Archiving System, a web archiving management system developed by the National Library of Australia http://www.nla.gov.au/nla/staffpaper/2004/koerbin2.html

PANDORA

Australia's Web Archive http://pandora.nla.gov.au/. The name, PANDORA, is an acronym that encapsulates their mission: Preserving and Accessing Networked Documentary Resources of Australia.

PDF

Portable Document Format. A proprietary but open standard file format for representing two dimensional documents.

PICT

Macintosh Quickdraw/PICT Drawing. An image file format with lossy compression.

PNG

Portable Network Graphics. An image file format with lossless compression.

PSD

Photoshop Format. An image file format.

PREMIS

Preservation Metadata Implementation Strategies http://www.loc.gov/standards/premis/. The Preservation Metadata: Implementation Strategies Working Group initially developed the PREMIS data dictionary as a specification with the goal of creating an implementable set of "core" preservation metadata elements, with broad applicability within the digital preservation community.

PRESTA

PREMIS Requirement Statement, a project of APSR http://www.apsr.edu.au/currentprojects/currentprojects.htm#presta. The objective of this

project is to develop a requirements specification for preservation metadata based on PREMIS.

RLG
Research Libraries Group.

RM
Real Media. An proprietary lossy compression streaming media file.

RTF
Rich Text Format.

RTSP
Real Time Streaming Protocol, a protocol that is used to deliver streaming media files.

SGML
Standard Generalized Markup Language.

SIP
Submission Information Package. A concept from the OAIS Reference Model.

SMPTE
Society of Motion Picture and Television Engineers http://www.smpte.org/

SVG
Scalable Vector Graphics. A language for describing two-dimensional graphics and graphical applications in XML.

TIFF
Tagged Image File Format. A file format for mainly storing images.

UQ
The University of Queensland. http://www.uq.edu.au/

US-ASCII
American Standard Code for Information Interchange. A character encoding based on the English alphabet.

UTF
Unicode Transformation Formats. Unicode is an industry standard designed to allow text and symbols from all of the writing systems of the world to be consistently represented and manipulated by computers.

WAV or WAVE
Short for Waveform audio format. An audio file format standard.

WMV
Windows Media File. A proprietary lossy compression streaming media file.

XML
Extensible Markup Language.

Appendix 10: Bibliography

Bradley, K. 2005, A*PSR Sustainability Issues Discussion Paper* , Available at http://www.apsr.edu.au/documents/APSR_Sustainability_Issues_Paper.pdf

Bradley, K. 2006, *Digital sustainability and digital repositories* , Available at http://www.valaconf.org/vala2006/papers2006/45_Bradley_Final.pdf

Bradley, K. & Henty, M. 2005, *Survey of Data Collections. Australian Partnership for Sustainable Repositories*, Available at http://www.apsr.edu.au/publications/data_collections.htm

Caplan, P. 2004, 'PREMIS - Preservation metadata - Implementation strategies update 1. Implementing preservation repositories for digital materials: current practice and emerging trends in the cultural heritage community', *RLG DigiNews*, [Online] Available at http://www.rlg.org/en/page.php?Page_ID=20462#article2

Consultative Committee for Space Data Systems 2002, *Recommendation for space data system standards. Reference Model for an Open Archival Information System (OAIS)*, Available at http://public.ccsds.org/publications/archive/650x0b1.pdf

Consultative Committee for Space Data Systems 2004, *Recommendation for space data system standards. Producer-Archive Interface Methodology Abstract Standard*, Available at http://public.ccsds.org/publications/archive/651x0b1.pdf

DROID, Available at http://www.nationalarchives.gov.uk/aboutapps/pronom/droid.htm

European Broadcasting Union, 2001, *Broadcast Wave Format - a format for audio data files in broadcasting*, Technical Specification 3285 Version 1.0, European Broadcasting Union, Geneva

FCLA Digital Archive 2005, *Digital Archive information* , Available at http://www.fcla.edu/digitalArchive/daInfo.htm

FILExt: The File Extension Source, Available at http://www.filext.com/

Hunter, J. & Choudhury, S 2005, *PANIC - an Integrated Approach to the Preservation of Complex Digital Objects using Semantic Web Services* , Available at http://www.iwaw.net/05/papers/iwaw05-hunter.pdf

IEEE LTSC CMI Working Group. *2005, The RAMLET project— Developing a reference model for resource aggregation for learning, education, and training* , Available at http://ieeeltsc.org/wg11CMI/ramlet/Pub/RAMLET_project_description.pdf

JHOVE - JSTOR/Harvard Object Validation Environment 2006, Available at http://hul.harvard.edu/jhove/

Library of Congress 2004, *AV Prototype Project Working Documents. Extension Schemas for the Metadata Encoding and Transmission Standard*, Available at http://www.loc.gov/rr/mopic/avprot/metsmenu2.html

Library of Congress 2005, *Digital Audio-Visual Preservation Prototyping Projects*, Available at http://www.loc.gov/rr/mopic/avprot/

Library of Congress, *Sustainability of Digital Formats, Planning for Library of Congress Collections*, Available at http://www.digitalpreservation.gov/formats/

METS: An Overview & Tutorial 2003, Available at http://www.loc.gov/standards/mets/METSOverview.html

Open Archives Initiative *2002, The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14*,
Current version available at http://www.openarchives.org/OAI/openarchivesprotocol.html

Langley, S. 2005, 'Where Are We?' in *Vital Signs: Creative Practice & New Media Now*, [Online] Available at http://search.informit.com.au/documentSummary;dn=031116469214359;res=E-LIBRARY

Lavoie, B. F. 2004, *Technology Watch Report. The Open Archival Information System Reference Model: Introductory Guide* , Available at http://www.dpconline.org/docs/lavoie_OAIS.pdf

Lavoie, B. & Gartner, R. 2005, *Technology Watch Report. Preservation Metadata* , Available at http://www.dpconline.org/docs/reports/dpctw05-01.pdf

National Library of Australia 2002, *Digital Preservation Policy*, Current version available at http://www.nla.gov.au/policy/digpres.html

National Library of New Zealand Metadata Extraction Tool Version 1.0 , Available at http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction

The OCLC/RLG Working Group on Preservation Metadata 2002, *Preservation Metadata and the OAIS Information Model. A Metadata Framework to Support the Preservation of Digital Objects* , Available at http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

The PREMIS Working Group 2004, *Implementing preservation repositories for digital Materials: current practice and emerging trends in the cultural heritage community* , Available at http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

PREMIS Working Group 2005, *Data Dictionary for Preservation Metadata*, Available at http://www.oclc.org/research/projects/pmwg/premis-final.pdf

RLG 2005, *An Audit Checklist for the Certification of Trusted Digital Repositories. Draft for Public Comment* , Available at http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Shirky, C. 2005, 'AIHT. Conceptual issues from practical tests' , in *D-Lib Magazine*, [Online], Available at http://www.dlib.org/dlib/december05/shirky/12shirky.html

Society of Motion Picture and Television Engineers, 2004, *SMPTE Metadata Dictionary*, Available at http://www.smpte-ra.org/mdd/index.html

Stanescu, A. 2004, 'Assessing the durability of formats in a digital preservation environment', *D-Lib Magazine*, [Online] Available at http://www.dlib.org/dlib/november04/stanescu/11stanescu.html

University of Leeds The Representation and Rendering Project, 2004, *Survey and assessment of sources of information on file formats and software documentation* , Available at http://ourweb.nla.gov.au/committees/apsr/prestapublicreport/www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf