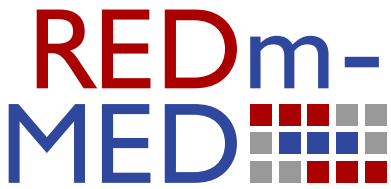

Minimum Mandatory Metadata Set for RAIDmap

Alex Ball

redm1rep111124ab10.pdf

Issue date: 15th June 2012

Catalogue Entry

Title	Minimum Mandatory Metadata Set for RAIDmap
Creator	Alex Ball (author)
Subject	data management; metadata
Description	This document defines a set of metadata elements corresponding to information considered necessary for good data management, most easily provided close to the point of record creation, and unique to a data record, data case or data development process. The set corresponds to a subset of elements drawn from the PREMIS and DataCite metadata schemata. It is provided as a basis for the metadata collection functionality of the RAIDmap Associative Tool.
Publisher	University of Bath
Date	24th November 2011 (creation)
Version	1.0
Type	Text
Format	Portable Document Format version 1.5
Resource Identifier	redm1rep111124ab10
Language	English
Rights	

Citation Guidelines

Alex Ball. (2012). *Minimum Mandatory Metadata Set for RAIDmap* (version 1.0). REDm-MED Project Document redm1rep111124ab10. Bath, UK: University of Bath.

CONTENTS

1 Introduction . . . 3
2 Theoretical basis . . . 3
3 Preservation metadata . . . 5
3.1 Object metadata . . . 5
3.2 Event metadata . . . 7
3.3 Agent metadata . . . 8
3.4 Rights metadata . . . 8
4 Descriptive metadata . . . 8
5 Minimum Mandatory Metadata Set . . . 10
5.1 Data cases . . . 10
5.2 Data records . . . 10
5.3 Data development processes . . . 11
6 Optional metadata . . . 11
References . . . 11

1 INTRODUCTION

It is critically important for the effective management of research data records that they are adequately documented. Such metadata enable the data to be understood, rendered, reused, repurposed, preserved, and properly controlled in terms of security and access. As a step towards this, the Research Activity Information Development Associative Tool (RAIDmap) will collect metadata about the data objects for which it records associations. In order to keep this collection activity to a manageable level, a Minimum Mandatory Metadata Set has been devised. This set contains the metadata that are (*a*) unique to the record in question, and (*b*) most efficiently collected at or before the point at which the record is added to the RAIDmap system. It should be understood that this set is not complete, in the sense of containing all the metadata necessary for good data management; the 'missing' metadata should be collected at other points in the curation lifecycle. This approach is inspired by that taken by the KIM Project [Bal06].

2 THEORETICAL BASIS

Over the past decade, a considerable amount of work has gone into enumerating the metadata required for good data management. One of the most influential contributions was the Open Archival Information System (OAIS) Reference Model [CCS02]. It was developed by the Consultative Committee for Space Data Systems as a standard vocabulary for describing the functions, processes, holdings and stakeholders of archival repositories for data. The OAIS Information Model explicitly mentions seven types of metadata that ought to be present:

- *Representation information* (needed to turn a bitstream into something meaningful);
- *Provenance information* (detailing the origin of the digital object and what has happened to it since);
- *Context information* (describing relationships and interactions with other entities);
- *Reference information* (providing identifiers for the digital object);
- *Fixity information* (for detecting or preventing changes to the digital object);
- *Packaging information* (for associating the above information with the target bitstream, thereby forming an Information Package)
- *Descriptive information* (for discovery services such as catalogues).

This information model was developed into two different practical metadata schemata by the CURL Exemplars in Digital Archives (CEDARS) project [Ced00], and the Networked European Deposit Library (NEDLIB) project [LM00]. A third, rather different schema was devised by the National Library of Australia (NLA) [NLA99], informed not only by the OAIS model but also NLA's PANDORA Project[1] and RLG's PRESERV specification [RLG98].

In 2001, the OCLC/RLG Preservation Metadata Framework Working Group published a report entitled *A Metadata Framework to Support the Preservation of Digital Objects* [OR02] which took the CEDARS, NEDLIB and NLA schemata (along with the OCLC Digital

1. PANDORA Project Web page, URL: `http://pandora.nla.gov.au/`.

Archive schema) and combined them to form a new schema, explicitly laid out according to the OAIS Information Model. This schema was only expressed in terms of a list of human-readable metadata elements, however, and lacked the formalism required for practical implementation. In order to address this point, the OCLC/RLG Preservation Metadata Implementation Strategies (PREMIS) Working Group was established in 2003. The group decided to take on board some innovations introduced into a version of the NLA schema developed by the National Library of New Zealand [NLN03], and reworked the previous working group's schema around a data model with five entities: intellectual entities (coherent sets of content), objects (digital realisations of content), events, agents and rights. The result was the PREMIS Data Dictionary, which provided a metadata schema formal enough to be directly implemented in machine readable format [PRE05]. Significant additions to PREMIS from the previous framework included explicit provisions for referencing file format registries and for digital signatures, and semantic units associated with the Agent entity. On the other hand, PREMIS was more streamlined with regards to recording software and hardware requirements and object characteristics.

As well as the new data model, the other innovation in PREMIS version 1.0 was a digital object typology (inspired by but dissimilar to the typologies in the Australasian schemas). The three categories of digital object were: file, bitstream (a complete stream of digital data, without the encapsulating data needed for it to exist as a file) and representation (the sum total of files making up a digital version of an intellectual entity).

After the conclusion of the work of the PREMIS Working Group, responsibility for the maintenance of the Data Dictionary moved to the Library of Congress. The Editorial Committee collected feedback from users of PREMIS over an eighteen-month period, then issued a major revision of the Data Dictionary in 2008 [PRE08], along with an official XML schema implementation. A further revision was issued in 2011 in response to feedback from a substantially increased user base [PRE11].

In some respects PREMIS represents the culmination of an entire branch of research into preservation metadata.[2] The aim of PREMIS was to produce a schema that encompassed the metadata that 'most working preservation repositories are likely to need to know in order to support digital preservation.' [PRE11, p. 3]. It is not exhaustive; it does not deal with format-specific technical metadata, nor does it specify detailed documentation of the hardware environment(s) known to support the data. Nevertheless, as the motivation for these simplifications resonate with those of REDm-MED, it is appropriate for the Project to consider PREMIS a complete set of preservation metadata, from which to draw elements for the Minimum Mandatory Metadata Set.

Being focused on preservation, PREMIS does not include packaging or descriptive information. In practice, considerations of packaging information are dealt with by choosing a well-supported packaging format such as the Metadata Encoding and Transmission Standard (METS), and such information is out of scope for the Minimum Mandatory Metadata Set.[3] There is no single standard enjoying widespread adoption for descriptive information concerning research data [Bal09]. The nearest scheme that exists to this is probably the DataCite Metadata Schema [Sta+11], which is used when datasets are registered with DataCite and assigned Digital Object Identifiers (DOIs). This schema will

2. For more detail on the background to PREMIS and other research into preservation metadata, see Caplan [Cap06].
3. METS Web site, URL: `http://www.loc.gov/standards/mets/`

therefore be used by REDm-MED as the target for a complete set of descriptive metadata, from which elements will be drawn for the Minimum Mandatory Metadata Set.

3 PRESERVATION METADATA

The following tables show how the Minimum Mandatory Metadata Set would contribute to a full preservation metadata record based on the PREMIS Data Dictionary Version 2.1.

3.1 Object metadata

In PREMIS, an Object is a file, a bitstream (encoded information within a file) or a representation (a complex object such as a Web page, made up of several files). The concept maps to that of data records in RAID.

PREMIS Semantic Unit	Application to RAIDmap
objectIdentifierType objectIdentifierValue	RAIDmap will assign its own unique identifier to each record it maps, and record it as **Identifier**.
objectCategory *(representation, file, bitstream)*	This will be encoded by the RAID modelling, rather than as textual metadata.
preservationLevel	How a file will be preserved is a matter for the holding repository.
significantProperties	The properties to be preserved should be agreed with the holding repository.
compositionLevel *(number of encryptions/compressions)*	Implicit: Objects will be modelled at a composition level where they can be rendered or edited by software tools.
fixity *(checksum type & value)*	Generated by automated tools at point of ingest into a repository. (Integrity checking of data records pre-ingest may be introduced into RAIDmap in a future phase, but will not be available in the initial release.)
size	Implicit: Derived by automated tools at point of ingest into a repository.
formatName	**File format** should be determined by RAIDmap, falling back to user-supplied information.
formatVersion	**File format version** should be determined by RAIDmap, falling back to user-supplied information
formatRegistryName formatRegistryKey formatRegistryRole	RAIDmap may use this alternative way of recording formats if available.

PREMIS Semantic Unit	Application to RAIDmap
formatNote *(e.g. 'tentative')*	Support for this may be added to RAIDmap if found useful.
creatingApplicationName	**Creating application** should be determined by RAIDmap, falling back to user-supplied information.
creatingApplicationVersion	**Creating application version** should be determined by RAIDmap, falling back to user-supplied information.
dateCreatedByApplication	**Date created** and **Date modified** should be determined by RAIDmap, falling back to user-supplied information
inhibitorType inhibitorTarget inhibitorKey	Optional: If features of the file have been password-protected or otherwise inhibited, RAIDmap should record the type of restriction, what is restricted, and the password (securely stored) as a **Technical restriction**.
originalName	Recorded as **Filename**.
contentLocationType contentLocationValue	The network location of the file will be recorded as **Location** and updated as necessary.
storageMedium	Implicit: This may be determined from the **Location**. Removable media should not be used as the primary location for storing records.
environmentCharacteristic environmentPurpose environmentNote	Implicit: RAIDmap will record the environment in which the data records were created. Other environments (for future editing or rendering) may be added later by a holding repository.
dependency	RAIDmap will record interdependencies between data records, but not necessarily as part of a data record's mandatory metadata.
software *(name, version, type, dependency)*	Optional: In most cases, **Creating application** is sufficient. Alternative suitable software may be monitored by a repository. If a specialist plugin was used by the **Creating application**, this should be recorded as a **Software dependency**.
hardware *(name, type)*	Optional: In most cases, the hardware requirements are known for the **Creating application**. In cases where specialist hardware is required, this should be recorded as **Hardware dependency**.

PREMIS Semantic Unit	Application to RAIDmap
signature	Generated by automated tools at point of ingest into a repository.
relationship	RAIDmap will record relationships between data records, but not necessarily as part of a data record's mandatory metadata.
linkingEventIdentifier	RAIDmap will record relationships between data records and data developmental processes, but not necessarily as part of a data record's mandatory metadata. Non-developmental events (e.g. fixity checking) would be recorded in the Data Management Plan rather than RAIDmap.
linkingIntellectualEntity-Identifier	RAIDmap will record relationships between data records and data cases, and between data records that differ only in format, but not necessarily as part of a data record's mandatory metadata.
linkingRightsStatement	**Rights** information will be recorded as one or more keywords; these keywords should be explained in the associated Data Management Plan.

3.2 Event metadata

Events in PREMIS relate to both development actions in RAID and non-developmental events such as fixity checking and back-ups.

PREMIS Semantic Unit	Application to RAIDmap
eventIdentifier	Artefact of the PREMIS data model.
eventType	Represented in RAID by the name of an action. For other events, this should be explicitly recorded in the Data Management Plan.
eventDateTime	Recorded as **Date and time** of an action.
eventDetail	Recorded in RAID as an annotation on an action.
eventOutcome eventOutcomeDetail	Applies to non-developmental events. Outcomes of, e.g. fixity checks will be assumed to be successful unless otherwise recorded in the Data Management Plan.
linkingAgentIdentifier	Recorded as **Agent** by RAIDmap. Should also be explicit for events recorded in the Data Management Plan.

PREMIS Semantic Unit	Application to RAIDmap
linkingObjectIdentifier	This will be encoded by the RAID modelling, rather than as textual metadata. Should also be explicit for events recorded in the Data Management Plan

3.3 Agent metadata

It is recommended that RAIDMap records agents using an identifier, and keeps a register mapping these identifiers to real names, contact details, etc. for display purposes. It may be beneficial for RAIDmap to distinguish between different types of agent (e.g. person, organization, software).

3.4 Rights metadata

The rights information recorded natively by PREMIS is mainly concerned with copyright, licences, and statutory exemptions and obligations. It also provides an extension point for expressing rights information in other ways.

For the REDm-MED Project, it will be important to know about

- *ownership of data* (whether the default position of University ownership applies, or some other contractual agreement applies);
- *ownership of rights* such as copyright, database right, moral rights and so on, where applicable;
- *licences* granted by the rights holders to users;
- *impediments to access* such as confidentiality agreements, ethical constraints and embargo periods;
- *obligations* placed on users and custodians of the data, e.g. retention periods.

This information should be provided in the Data Management Plan. To support cases where the rights situation varies between records, the mandatory metadata element **Rights** is provided. This element should contain a keyword or keywords, which can be used to look up the relevant rights information in the Data Management Plan. Exactly one set of rights information in the Data Management Plan should be indicated as the default; a keyword should be provided for referring to that default.

4 DESCRIPTIVE METADATA

The following table shows how the Minimum Mandatory Metadata Set would contribute to a full metadata record based on the DataCite Metadata Schema Version 2.2.

PREMIS Semantic Unit	Application to RAIDmap
Identifier	Recorded as **Identifier**.
Creator	Recorded as **Creator**, defaulting to the **Agent** associated with the first action associated with the record.
Title	**Title** should be determined by RAIDmap, falling back to user-supplied information.
Publisher	Not applicable until published.
PublicationYear	Not applicable until published.
Subject	Recorded at the data case level as **Subject**, possibly pre-populated via the associated **Project**.
Contributor	Implicit: This can be generated from the **Agent**s associated with the actions leading into the data record.
Date	DataCite defines several date types, such as 'Created', 'Updated', and 'Valid'. RAIDmap will record **Date created** and **Date modified**.
Language	Optional: Data records will be assumed to be in (British) English unless otherwise indicated by **Language**.
ResourceType	**Type** (e.g. spreadsheet, graph, model) should be determined by RAIDmap, falling back to user-supplied information. DataCite's resourceTypeGeneral may be inferred from the given **Type**.
AlternateIdentifier	This would be of limited use prior to ingest into a repository.
RelatedIdentifier	This will be encoded by the RAID modelling, rather than as textual metadata.
Size	Implicit: Derived by automated tools at point of ingest into a repository.
Format	**File format** and **File format version** should be determined by RAIDmap, falling back to user-supplied information.
Version	Recorded as **Version**.
Rights	See subsection 3.4.
Description	A brief (one-line) description of the record should be recorded as **Description**.

5 MINIMUM MANDATORY METADATA SET

The following tables lists the mandatory metadata elements for data cases, data records and data development processes respectively, and describe how each metadata element will be collected.

5.1 Data cases

Metadata element	Collection method
Project	Supplied by the user from suggestion list.
Subject	Supplied by the user from suggestion list.

5.2 Data records

Metadata element	Collection method
Identifer	RAIDmap will generate a persistent and unique identifier.
Title	Determined by RAIDmap, falling back to user entry.
Version	Determined by RAIDmap, falling back to user entry.
Description	Supplied by the user.
Type	Inferred by RAIDmap, corrected if necessary by the user.
File format *(name and version)*	Determined by RAIDmap, corrected if necessary by the user.
Creating application *(name and version)*	Determined or inferred by RAIDmap, corrected if necessary by the user.
Date created	Determined by RAIDmap, corrected if necessary by the user.
Date modified	Determined by RAIDmap, corrected if necessary by the user.
Creator	Determined by RAIDmap, corrected if necessary by the user.
Owner *(day-to-day custodian)*	Inferred by RAIDmap, corrected if necessary by the user.
Responsible entity *(DM overseer)*	Inferred by RAIDmap, corrected if necessary by the user.

Metadata element	Collection method
Rights holder	Default provided by RAIDmap, corrected if necessary by the user.
Rights	Default provided by RAIDmap, corrected if necessary by the user.
Filename	Determined by RAIDmap.
Location	Determined by RAIDmap, corrected if necessary by the user.

5.3 Data development processes

Metadata element	Collection method
Date and time	User entry, but default value supplied by RAIDmap.
Agent	User entry, but default value supplied by RAIDmap.

6 OPTIONAL METADATA

The following table lists optional metadata that may be collected for data records.

Metadata element	Collection method
File size	Determined by RAIDmap.
Software dependency *(plug-in, add-on)*	Supplied by the user.
Hardware dependency	Supplied by the user.
Technical restriction *(type of restriction, subject of restriction, password)*	Determined as far as possible by RAIDmap, corrected and extended if necessary by the user.
Language	Determined by RAIDmap, corrected if necessary by the user.

REFERENCES

[Bal06] A Ball (2006-08-10). *The KIM Minimum Mandatory Metadata Set.* KIM Communication kim41com004ab10. University of Bath.

[Bal09] A Ball (2009-06). *Scientific Data Application Profile Scoping Study Report.* UKOLN, University of Bath. URL: `http://www.ukoln.ac.uk/projects/sdapss/papers/ball2009sda-v11.pdf`.

[Cap06] P Caplan (2006-07). 'Preservation Metadata'. In: *DCC Digital Curation Manual*. Ed. by S Ross & M Day. Digital Curation Centre: Edinburgh. URL: http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata/.

[CCS02] Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. Blue Book CCSDS 650.0-B-1. Also published as ISO 14721:2003. URL: http://public.ccsds.org/publications/archive/650x0b1.pdf.

[Ced00] Cedars Project (2000-03). *Metadata for Digital Preservation: the Cedars Project Outline Specification*. University of Leeds. URL: http://www.webarchive.org.uk/wayback/archive/20050410120000/http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html.

[LM00] C Lupovici & J Masanès (2000-09). *Metadata for the Long-Term Preservation of Electronic Publications*. NEDLIB Report 2. Koninklijke Bibliotheek: The Hague. URL: http://www.kb.nl/hrd/dd/dd_links_en_publicaties/nedlib/NEDLIBmetadata.pdf.

[NLA99] National Library of Australia (1999-10). *Preservation Metadata for Digital Collections*. URL: http://www.nla.gov.au/preserve/pmeta.html.

[NLN03] National Library of New Zealand (2003-06). *Metadata Standards Framework: Preservation Metadata (Revised)*. URL: http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf.

[OR02] OCLC/RLG Working Group on Preservation Metadata (2002-06). *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*. OCLC: Dublin, OH. URL: http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.

[PRE05] PREMIS Working Group (2005-05). *Data Dictionary for Preservation Metadata*. Final Report. OCLC & RLG: Dublin, OH & Mountain View, CA. URL: http://www.oclc.org/research/projects/pmwg/premis-final.pdf.

[PRE08] PREMIS Editorial Committee (2008-03). *PREMIS Data Dictionary for Preservation Metadata*. Version 2.0. Library of Congress: Washington, DC. URL: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.

[PRE11] PREMIS Editorial Committee (2011-01). *PREMIS Data Dictionary for Preservation Metadata*. Version 2.1. Library of Congress: Washington, DC. URL: http://www.loc.gov/standards/premis/v2/premis-2-1.pdf.

[RLG98] RLG Working Group on Preservation Issues of Metadata (1998-05). *Recommended Preservation Metadata Elements for Digital Master Files*. Final Report. RLG: Mountain View, CA. URL: http://www.oclc.org/research/activities/past/rlg/digpresmetadata/report.htm.

[Sta+11] J Starr et al. (2011-07). *DataCite Metadata Schema for the Publication and Citation of Research Data*. Version 2.2. DataCite Consortium. DOI: 10.5438/0005.