

RUBRIC Toolkit: Metadata

Table of Contents

What are the purposes of metadata?	2
What metadata enhances the repository's interoperability?	2
What metadata facilitates resource discovery?	3
What metadata enhances preservation and authentication?	4
Explanation of terms.....	4
How to choose a metadata schema for resource discovery	5
Criteria for deciding what metadata schema to use	6
How many schema are needed?	6
What metadata schema meet the above criteria?	6
Other schema.....	7
Metadata for images and videos	11
Metadata for digital theses	11
How to ensure your repository will be harvested by service providers	12
Harvesting explained.....	12
Harvesting 1: Metadata for harvesting.....	13
Harvesting 2: Connecting with the harvesters.....	16
Harvesting 3: Australasian Digital Thesis.....	18
Metadata management and quality control	21
Self-submission and metadata.....	21
Quality control.....	21
Entering Metadata: Guides and Tools for Repositories	22

Further Guides to Metadata.....	22
References and Further Reading.....	23

What Are the Purposes of Metadata?

Metadata is information that is provided about or alongside a resource to inform users what that resource is about, what conditions govern its use, the kind of format of the resource, and more. Metadata is used for the following purposes:

- **descriptive** metadata enables resource discovery by providing subject and other descriptions of the resource (e.g. title, author, abstract)
- **rights** metadata informs users of the legal (copyright) conditions of access to the resource
- **structural** metadata enables the correct technical storage and display format of resource files
- **preservation** metadata facilitates long term maintenance and preservation of resources by recording technical specifications about the file type and requirements for its use
- **versioning** metadata maintains the authenticity and integrity of the repository by monitoring each modification made to its records and files
- **administrative** metadata for in-house administration of the repository.

What Metadata Enhances the Repository's Interoperability?

Repositories need to be interoperable with various internet harvesters in order to understand their search requests and return them meaningful information. They must be able to migrate records from other databases and transfer their archive to another repository when an upgrade is required.

There are different metadata schemas, each designed for specific types of resources and institutional purposes. The most widely known schema that aims to be a minimal catch-all for all metadata and providing maximum internet interoperability, is the [Dublin Core](#)¹ schema, or more specifically the Unqualified or Simple Dublin Core schema. This particular schema is explained in further detail below.

Different repositories also use more granular schemes to make it relatively easy to transfer data to other widely supported schemas. Some of those used are MARC in its XML format (MARCXML), and Qualified Dublin Core, which is more granular than the basic 15 elements available to Unqualified DC. Data stored in these schemas is in turn can be mapped to future systems. Clearly the more granular the storage of the original data the greater the

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

potential for its preservation in the format and for the breakdown of detail if desired.

What metadata is required to ensure interoperability?

Mandatory: Simple Dublin Core

For future migrations: Widely supported and granular schema (e.g. MARCXML)

What Metadata Facilitates Resource Discovery?

“Discoverability” is inseparable from “interoperability” within the internet world of searching and harvesting. The first priority for being discovered is to have data that can be expressed within a Simple Dublin Core schema. For example:

- <dc.title> Metadata in Practice
- <dc.creator> Hillmann, Diane I.
- <dc.date> 2004

Details on how to enter data to ensure discoverability via Dublin Core can be found in: [Entering Metadata: Guides and Tools for Repositories](#).

While simple or unqualified DC is the minimum requirement for resource discovery, other standard schemas are used in addition to unqualified DC for more finely grained communication among specialist communities and for specialist resource types:

- libraries are used to communicating through [MARC](#)², which can be adapted for repositories in XML format, MARCXML. Some repositories store data in MARCXML format
- government departments have developed their own schemas or locator services e.g. [AGLS](#)³ (Australia); [NZGLS](#)⁴ (New Zealand); [GILS](#)⁵ United States
- a separate metadata schema for specialist archival collections, the [EAD](#)⁶ (Encoded Archival Description), is also widely used for specialist archival databases
- [VRA](#)⁷ (Visual Resources Association) is widely used for images and visual works. [MIX](#)⁸ is another. Others for art resources are listed at the [Getty](#)⁹ site
- [TEI](#)¹⁰ (Text Encoding Initiative) was designed for Humanities scholars to assist with the encoding of different views of large text documents
- thesis and dissertation metadata schema include [ETD-MS](#)¹¹, [UKETD_DC](#)¹², [TDL](#)¹³, [XmetaDiss](#)¹⁴ (German), [TEF](#)¹⁵ (French)
- [Qualified Dublin Core](#)¹⁶ is another more generic schema which can be used to enhance discovery through more granular storage of different types of data values
- [MODS](#)¹⁷ (Metadata Object Description Schema) is an evolutionary development of MARC. MODS uses simplified MARC tags and re-written them in natural language

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

XML format. MODS is designed to carry library catalogue quality data into the harvestable world of wider internet discovery.

What metadata enhances resource discovery?

Mandatory: Simple Dublin Core

Recommended: More granular widely recognised standard schema

Recommended: Widely supported schema for specialist communities and for specialist resource types

What Metadata Enhances Preservation and Authentication?

Much of the required metadata for preservation of the archive and monitoring of changes made to its records with their attached document files will be machine generated at ingest. But repository managers need to be aware of how extensive their particular repository is in these areas.

[JHOVE](#)¹⁸ metadata files identify the format (e.g. pdf, jpeg) of an ingested file, its validation (i.e. verification that it is well-formed) and its technical characteristics.

[PREMIS](#)¹⁹ metadata is currently being developed by a working group sponsored by [OCLC](#)²⁰ and [RLG](#)²¹. Its aim is to establish comprehensive preservation metadata covering:

- provenance: who has had custody/ownership of the digital object?
- authenticity: is the digital object what it purports to be? (includes versioning audits)
- preservation activity: what has been done to preserve the digital object?
- technical environment: what is needed to render and use the digital object?
- rights management: what intellectual property rights must be observed?

Explanation of Terms

Following is an explanation of terms used throughout this section. Since there is no uniform standard of terminology in this field, the first two terms listed here may be used differently in other publications.

The following terms reflect those used in the [Open Archives Initiative Protocol for Metadata Harvesting](#)²² usage:

Item: a container or package that includes:

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

- a metadata record
- the archived document or paper to which the metadata refers
- and other supporting files and data that serve to preserve, monitor and assist searching of either the metadata record or the archived document.

The large archived document or paper may be broken up into multiple files and still be part of the single repository item.

Sometimes an item is called an object. In this context, a Learning Object Repository or a Digital Object Repository would be a repository containing “objects”, each object consisting of multiple datastreams as above. The preferred term used in this section is “item” rather than “object” because the word “object” is used in another sense (as explained under the term “resource” below).

Resource: The document or paper or article or thesis or image etc that is deposited in the repository will be referred to as the “resource”. It is the deposited document for which the metadata is created to describe it or assist with searching and preserving it, and managing its rights of access.

Resource in this context includes a “document-like object”. The [Open Archives Forum tutorial](#)²³ defines a document-like object as a digital data unit that is comparable to a paper document. The same tutorial defines a resource as an object the metadata is “about”. In this section of the Toolkit, the meaning of resource is restricted to the above definition.

Service provider: A service provider in the repository context is an internet harvester that searches and gathers for the display of repository records. An Australian national service provider is the National Library of Australia's [ARROW Discovery Service](#)²⁴; a major international service provider is [OAIster](#)²⁵.

Data provider: In this section, a data provider is a software program within the repository that provides data in an appropriate schema to a service provider. Data provider can also refer more generally to the repository or repository institution itself.

How to Choose a Metadata Schema for Resource Discovery

A recent UKOLN article - [Choosing a Metadata Standard For Research Discovery](#)²⁶ - outlines a checklist of guidelines to assist organisations in choosing a metadata schema for a repository.

A similar checklist was developed by Jeffrey Beall - [Metadata Schemes Points of Comparison](#)²⁷

These checklists can be conflated to five areas for consideration.

Criteria for Deciding What Metadata Schema to Use

The metadata needed for a repository must be compatible with the requirements of:

- **Interoperability:** the metadata must be compatible or migratable across different repository systems (the repository solution being used now will probably change); it must be searchable by major internet harvesters and search engines; it must be able to import from and export to other databases.
- **Extensibility and growth:** the metadata must be in a format that allows for future developments and expansion. Although the pilot repository may begin with the archiving of predominantly simple text formatted material, future requirements may well involve more complex resource types which will require 'extended' metadata. It is important to ensure that the base metadata records will allow for future extension with new resource types, and future enhanced versions of repository software and harvesting functions. Certain metadata schema may also have the capability of expanding to meet new requirements.
- **Sustainability:** the metadata schemas used, as well as the content and ways of entering data, will need to be measured against procedures, standards, technologies and reputable support infrastructure that are sustainable for the long-term.
- **Granularity:** generally, the more granular the data is in both content and format, the greater the capacity for both future developments and the ability to meet the requirements of the above criteria.
- **Ease of use and existing skills:** notwithstanding these considerations, it is obviously preferable to deploy metadata that is not overly complex to use and that makes use of the existing experience and skills of current staff.

How Many Schema Are Needed?

Because different metadata schemas are designed for different audiences and resource types, a number of metadata schemas should be considered, not just one. This is not as onerous as it sounds, since if a highly granular schema is used as the base schema for data entry and storage, then the repository can be configured to use that base metadata to create other metadata datastreams with different schema as required. Thus metadata can be stored in a MARC format, and from this MARCXML datastream a Simple Dublin Core file can be created when requested by a harvester.

What Metadata Schemas Meet the Above Criteria?

The baseline metadata for harvesting and interoperability is the Simple Dublin Core schema. The Open Archive Initiative (OAI) is an internationally recognised committee supporting interoperability standards, and it has developed the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). The OAI-PMH requires repositories to use the metadata

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

schema known as simple or unqualified Dublin Core. (Simple DC is sometimes referred to as OAI Dublin Core.) All OAI compliant harvesters look for this basic metadata. Some will look for other metadata as well a Dublin Core.

Simple Dublin Core

[Simple Dublin Core](#)²⁸ is a bare-bones metadata schema of up to 15 elements:

contributor	coverage	creator	date
description	format	identifier	language
publisher	relation	rights	source
subject	title	type	

The elements are repeatable so more than one author and subject can be assigned to a record. It is not necessary to include data for all 15 elements, so a technically valid (though not very useful) DC datastream may consist of only one element.

Repository comparison
 The VITAL, Fez, DSpace and EPrints repositories automatically generate a simple Dublin Core metadata datastream from the data that is entered by the depositors and editors of records.
 Data can be stored in these repositories in other metadata schema such as Qualified Dublin Core, MODS or a MARCXML, but the data provider program in the repository will generate the Simple Dublin Core from these more granular schema. There is no need for an editor to create a separate Dublin Core record manually.

Simple Dublin Core meets the criteria of interoperability since it has become the standard adopted by the [Open Archives Initiative Protocol for Metadata Harvesting \(OAI-PMH\)](#)²⁹. Dublin Core is international in scope and used across multiple languages.

Although it lacks granularity, Dublin Core was designed to be used in conjunction with more granular metadata schemas. It can be generated from most other major metadata such as MARC, MODS and Qualified Dublin Core. It is thus compatible with the criteria of extensibility and growth. The active support infrastructure behind Dublin Core encourages confidence that it will be kept abreast of harvesting needs in the future.

Dublin Core consists of readily understood semantics that are few in number. Once set up and configured in the repository it is machine generated so it presents no difficulty in its use. It is also sustained by the [Dublin Core Metadata Initiative \(DCMI\)](#)³⁰ in collaboration with other major support institutions such as the [Library of Congress](#)³¹ and the [Open Archive Initiative \(OAI\)](#)³².

Other Schemas

In libraries, the most time-tested descriptive metadata for users is MARC. MODS has been developed particularly for library applications from MARC as a natural language tag XML metadata schema. These have been used in both the Fez and VITAL repositories. DSpace uses Qualified Dublin Core as its primary metadata schema.

MARC

Libraries are familiar with metadata in the MACHine Readable Cataloging ([MARC](#)³³) format and this has proven its worth since its widespread adoption. Its scope allows for a focused breakdown of different types of subjects and descriptions based on both controlled and free-text vocabularies. It also contains multiple fields and subfields for rights metadata. In its XML format it is known as MARCXML.

The NSDL [OAI Best Practices](#)³⁴ website suggests that “MARCXML may be a good option for an additional metadata schema to expose via OAI for data providers who:

- locally describe resources in MARC according to [AACR2r](#)³⁵, and
- have as a primary audience for resources described via OAI records the core library community.

Repository comparison

The VITAL repository supports MARC data in XML format. The [SIMILE Project](#)³⁶ is planning to improve the support DSpace can offer for different metadata schemas and extensions including MARCXML. Fez has successfully used MARC but is now using MODS, an XML natural language derivative of MARC.

While much detail can be stored in a MARCXML record, there could well be a problem with information overload if all the data displayed with an item's record in the repository portal. The repository can be configured to display only certain fields on different pages within the repository. Thus an item's main title page may display title, author and abstract, but the subjects and keywords linked to that item may only appear in a Browse or Search by Subject page.

MARCXML meets the criteria of interoperability largely as a result of its granularity. Its granularity is also a reason it is widely seen as surviving for some time yet into the future. Its storage of detailed information in subfields within other fields makes it an excellent storage of data that can be drawn on for use by other metadata schema. To this extent it is an excellent support for the extensibility needs of future metadata and the growth and evolution of repositories.

It has also proven itself sustainable with the support of the Library of Congress. However, it is not considered easy to use, although this is offset in many libraries by the presence of trained cataloguers skilled in its use.

MARC was not originally designed for the current XML based technology (although MARC can be used in XML format) and for that reason the Library of Congress has prepared a new

metadata schema, MODS, that is more compatible with the current technology and metadata schemas. VITAL 3.0 supports MODS and other metadata schemas as well as MARC.

MODS

The Metadata Object Descriptive Schema ([MODS](#)³⁷) has been developed by the Library of Congress as an XML schema that is largely based on MARC and is designed to rationalise some of the fields of MARC and employ natural language tags. It is intended particularly as a library application.

The NSDL [OAI Best Practices](#)³⁸ website suggests that “MODS may be a good option for an additional metadata schema to expose via OAI for data providers who:

- locally engage in descriptive practices heavily influenced by resource description standards in libraries, and
- have as a primary audience for resources described via OAI records a community well-versed in library descriptive practices, yet also want robust records in a format accessible to service providers outside the core library community.

Repository comparison

The Fez repository uses MODS and the basic VITAL 3.0 repository includes MODS as one of its configurations and transformations (along with MARC, EAD and TEI). RUBRIC is primarily using MODS with its testing and configurations of VITAL 3.0. The [SIMILE Project](#)³⁹ is planning to improve the support DSpace can offer for different metadata schemas and extensions including MARC.

One of the biggest advantages of including a MODS schema is that an author's name variations and affiliations can be nested in discrete containers that maintains their association with each authority or main name entry (these are lost in a flat metadata schema like Dublin Core). Thus:

```
<name type="personal">
  <namePart type="given">First name</namePart>
  <namePart type="family">Family name</namePart>
  <displayForm>Nonstandard form of name as it appears on the
resource</displayForm>
  <affiliation>Affiliated institution</affiliation>
</name>
```

MODS maps easily to Simple Dublin Core. While some granularity of MARC is lost in MODS, what is lost is largely in fields that have no consequence for repository purposes and archives.

For example, MARC language codes are found in 008/35-37 and 041 \$a for text material, 041 \$d for sung or spoken material and 041 \$e for librettos.

MODS conflates all of these into a single, but repeatable, <language><languageTerm> element.

The MARC 546 language note (free text) is mapped to the repeatable MODS <note> element, although this <note> can be further specified as a “language” note. Since the primary reason for including the language information is for the benefit of international harvesters, and since these harvesters generally look for standard codes for languages (e.g. rfc3066, iso639-2b), the MODS conflation of the multiple MARC language fields is an advantage for repositories.

Another example of the sort of granularity lost when MARC maps to MODS is in the uniform titles. MARC tags 130 for a main entry uniform title, 240 for a uniform title and 730 for an added entry uniform title, all map to MODS <title><titleInfo>type=“uniform”.

MODS meets the criteria of interoperability. It is an XML code that is increasingly recognised by service providers (harvesters).

MODS was designed as a library application to be the successor to MARC. Although it falls short of MARC's granularity, it maintains one of the highest levels of granularity of any metadata schema currently available.

This puts it in good stead for meeting extensibility and growth requirements. Being supported by the Library of Congress', ['Network Development & MARC Standards Office'⁴⁰](#), it also meets the criteria of sustainability.

With its natural language tags it is easier to use than MARC. It also uses fewer fields overall than MARC, rationalising several overlaps in MARC as demonstrated in previous examples.

QUALIFIED DUBLIN CORE

[Qualified Dublin Core⁴¹](#) is the base metadata schema used by the DSpace repository. Qualified Dublin Core extends the 15 elements of Simple Dublin Core by adding (a) element refinements to each element to make its meaning more specific, and (b) encoding schemes that consist of controlled vocabularies in order to interpret the element value. For example, the simple DC element “title” can have the element refinement “alternative” added to it. The element “identifier” can have the encoding scheme “URI” added to it to indicate the data entered in the identifier field is a web link.

The NSDL [OAI Best Practices⁴²](#) website suggests that “Qualified Dublin Core may be a good option for an additional metadata schema to expose via OAI for data providers who:

- have a need for more granularity of description than is available in simple Dublin Core but not a fundamentally different approach to resource description, and
- use controlled vocabularies that they wish to specify within their metadata records, and
- have resources of interest to many different knowledge communities with disparate descriptive metadata practices.

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

The base schema for DSpace

The DSpace repository contains a Dublin Core metadata registry that allows the editor to create qualified DC elements as required. Obviously this should as a rule comply with the element refinements and encoding schemes supported by [DCMI](#)⁴³, but an advantage of the qualified DC is that additional in-house element labels can be created for local use without compromising the value of the rest of the Dublin Core entries. (or can create a completely separate local metadata schema for in-house purposes.) As with Simple Dublin Core, all fields are repeatable.

Qualified Dublin Core also meets the standards of interoperability. It is widely recognised as a standard along with Simple Dublin Core by OAI harvesters. It's granularity is assisted by a [Library Application Profile](#)⁴⁴ qualified DC schema developed and supported by the [DCMI](#)⁴⁵. It can also be adapted for local in-house use without affecting the value of the DCMI supported elements. It has thus demonstrated a capacity for extensibility. It is a sustainable scheme given its support from [DCMI](#)⁴⁶ and is relatively easy to use.

Other

There are many other metadata schemas for various disciplines and resource types. A local institution is also free to create its own local metadata schema. All of these metadata schema have a place, and it may be desirable to use an "application" of more than one metadata schema. But for the purposes of establishing a Pilot Repository the schema most likely to be encountered and used at the beginning are covered here.

What Metadata Should be Used for Images and Videos

It is anticipated that a repository will be principally archiving text documents: journal article publications, conference and working papers, theses, book chapters. The above metadata schemas are designed to handle these and other resource types. Other resource types such as images and video material present additional preservation, sustainability and rights issues. Some of this material is measured in gigabytes and such size presents additional technical considerations. Some repositories may have difficulties handling files in excess of 10 or 20 megabytes. Some files rely on proprietary software that may have a limited lifespan.

MARC, MODS and Qualified Dublin Core have all been used for video and image resources. [Picture Australia](#)⁴⁷, the Australian National Library's online image collection, uses Qualified Dublin Core and a Simple DC datastream is created for harvest. MARC has long been used for storing image and video metadata and MODS is an XML development from MARC. The [MODS Implementation Registry](#)⁴⁸ lists MODS pilots with video and image metadata.

RUBRIC has prepared templates for basic image records that are suitable in cases where little more than an image format and size need be recorded. Art and special heritage types of images will normally require more detail.

Refer to the list under the above heading , 'What metadata facilitates resource discovery?' for

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

other image metadata schema.

Metadata for Digital Theses

Digital theses have specialist properties that have generated several specialist metadata schema for these. A list of some of the more widely used ones is under the above heading, 'What metadata facilitates resource discovery?' A detailed explanation for the requirements for thesis for the Australasian Digital Thesis Program is outlined below. Further discussion about the specialist properties for digital thesis metadata more generally can be found on the RUBRIC Metadata Specialist's personal blog, Metalogger: [Recommended Australian digital thesis metadata](#)⁴⁹.

A common query is whether there are standard terms that can be used throughout Australia for the many types of theses. While the generic term `thesis` will be adequate for most OAI harvesting requirements, where more granular descriptions are needed for reasons other than OAI harvesting, RUBRIC recommends terms used in [The Australian Qualification Framework Implementation Handbook](#)⁵⁰. The [Australian Qualifications Framework](#)⁵¹ would appear to offer the closest set of `standard` terms nation-wide:

- Doctoral Degree
 - research doctorate (*also an ADT thesis*)
 - professional doctorate
 - higher doctorate (*also an ADT thesis*)
- Masters Degree
 - coursework masters
 - research masters (*also an ADT thesis*)
 - professional (coursework) masters
- Graduate Diploma
- Graduate Certificate
- Bachelor Degree
 - there is a range of bachelor degree programs see the handbook, page 50
 - Bachelor Honours degree (use only this one as a Bachelor Degree sub-type)
- Advanced Diploma
- Diploma

How to Ensure Your Repository will be Harvested by Service Providers

Harvesting Explained

[The case for metadata harvesting](#)⁵² by (Simeoni 2003) is a useful article for an understanding of the importance of harvesting and the role it plays in federating the content of independent institutional repositories.

Harvesting:

- is an increasingly popular model of interaction between independent organisations for the purposes of resource discovery
- is based on the exchange of consistent metadata with a service provider
- can be inbuilt from the inception of the service
- builds on the interoperability of the repository software

Harvesting 1: Metadata for harvesting

The [ARROW Discovery Service](#)⁵³ has produced a [Harvesting Guide](#)⁵⁴ that recommends different levels of metadata content for harvesting. New Zealand's National Research Discovery Service (NRDS) is also currently reviewing different levels of metadata requirement for harvesting. Below are the current “metadata for harvesting” recommendations.

What is the ARROW Discovery Service?

The ARROW Discovery Service is an OAI-PMH compliant national service provider, or harvester, that is managed by the National Library of Australia. It liaises with Google, Google Scholar and Yahoo!, as well as with other service providers such as SCIRUS (Elsevier), NERIUS (European), EdNA Online (Australian), PerX (Scotland). These contacts enable the NLA/ARROW Discovery Service to negotiate the harvesting and harvesting conditions of Australian material by these search engines and providers. Institutional repository managers may liaise independently with the same providers or they may choose to register with the ARROW Discovery Service which can negotiate on their behalf.

The following elements are recommended as the minimum for harvestable metadata for any OAI-PMH service provider. A repository should be configured so that for each record metadata is entered and mapped to the Dublin Core elements below:

Dublin Core element	ARROW DS	NRDS (under review)
---------------------	----------	---------------------

title	Mandatory for discovery	Yes
subject	Optional for discovery	Yes
description	Optional for discovery	Yes
type	Mandatory for increased discovery and citations	Yes
source		(Yes)
relation	Optional for increased discovery and citations	(Yes)
coverage	Optional for international exchange	(Yes)
creator	Mandatory for discovery	Yes
publisher	Optional for increased discovery and citations	Yes
contributor	Optional	Yes
rights	Optional for discovery	Yes
date	Mandatory for discovery	Yes
format		(Yes)
identifier	Mandatory for discovery	Yes
language	Optional for international exchange	Yes

(Yes) in the NRDS column indicates Low Importance, but note that NRDS values are currently under review.

Title refers to the main title. Alternative titles can optionally be added in multiple title fields.

Subject refers principally to the Marsden codes in NRDS, but also (with lesser importance) to keywords and controlled vocabulary subjects. RFLD codes can also be configured for use in VITAL, Fez, DSpace and Eprints. There is scope for other subject entries such as keywords.

Description refers to the abstract or summary of the deposited resource. Note that this Dublin Core element is not restricted to an abstract as defined by MARC or academia, but can include a summary of any kind. Thus some library repositories have opted to ensure this field is always populated even if only with a summary entered by an editor and taken from a key section of the article if that is the only way to populate this field. One reason some do this is to overcome display issues in some harvesters (e.g. OAIster) in cases where this field is left blank.

Type refers to the genre of the resource deposited: e.g. journal article, conference paper, thesis. A controlled thesaurus should be used for this. Unfortunately there is at present no national or international standard thesaurus for this purpose. Although all Dublin Core fields

are technically repeatable, it is best that the “type” field appear only once. The ARROW Discovery Service harvester recommends that there be only resource type in the Simple Dublin Core field.

Source, according to the [DCMI element definitions](#)⁵⁵, refers to a resource that is related intellectually to the archived resource but does not fit easily into a “relation” element. (e.g. Source="Image from page 54 of the 1922 edition of Romeo and Juliet").

Relation refers to parent publication, such as the title and issue of the journal that published the journal article archived in the repository. It can also refer to a related resource, such as a resource that is located outside the repository. If the repository points to the URI of a resource that is at, say, its publisher's site and not in the repository, the URI for that resource should go in dc.relation. See the notes for Identifier.

Coverage refers to the geographical area and/or temporal period covered by the topic of the document archived.

Creator is the author.

Publisher is self-explanatory.

Contributor is sometimes used in place of creator for all authors of a document. (e.g. default in DSpace)

Rights covers information about copyright conditions of the resource, including access conditions. The field is repeatable.

Date refers generally to the date the resource was published or made public. It may also refer in the case of some document types (e.g. preprints) to the date the resource was created. There can be multiple date fields. Qualified Dublin Core can be used to explain the date type in each instance of a date value.

Format refers to the MIME type of the resource. This is normally a machine generated value.

Identifier can refer to a DOI, an ISBN, a Handle, a URL, a unique number in a working paper series, etc. A repository should be configured so that it can resolve a DOI, Handle or URL to take the user directly to the intended target. For purposes of OAI-PMH harvesting the identifier with the resolvable (URI) link should point to the repository's metadata page that describes and links to the resource. Other non-resolvable identifiers (e.g. an ISBN, a working paper unique number) will point to the resource itself just as all other DC elements refer to the resource. A resolvable link to a publisher's online version of the resource should be mapped to dc.relation, not dc.identifier, for the OAI Dublin Core datastream. Otherwise the OAI service provider will direct users away from the repository and directly to the resource. (See the OAI-PMH explanation of [Unique Identifier](#). This OAI identifier will normally be machine generated in an OAI compliant repository.)

Language should be configured in the repository so that it generates a standard ISO 639-2 or RFC 1766 code in the Dublin Core field.

Not all fifteen elements are always used by all metadata harvesters:

The **OAIster** service provider does not look for **relation** or **coverage**;
ARROW Discovery Service does not scan for **format** or **source**.

Minimal Metadata Essentials for Harvesting

- **title**
- **subject** (keywords or controlled vocabulary)
- **description** (abstract, other)
- **type*** (e.g. journal article, conference paper)
- **creator/contributor**
- **date**
- **identifier** (include URI where possible)

* do not repeat the “type” element

Harvesting 2: Connecting with the Harvesters

Linking up to Google and Yahoo!

The purpose of harvesting your repository content by a search engine is to improve its discoverability. Most users will only look at the first few pages of a search result to find the information they are looking for, so it is advantageous to have your search result display early on the list to increase its chance of being discovered. The location of the item from your repository in the list of search results is called its **ranking**.

Registering with harvesting services such as the [ARROW Discovery Service](#)⁵⁶ will give Australian university repositories a higher ranking in Google and Yahoo! results. The ARROW Discovery Service has negotiated directly with Google to secure higher rankings for its university repository harvested resources. It has also negotiated with a major U.S. harvester, [OAIster](#)⁵⁷, to secure rankings with the Yahoo! search engine. National Library of Australia's GATEWAYS journal has an [article](#)⁵⁸ that discussing how the ARROW Discovery Service achieves this end through agreements with Google and the OAIster service.

How can university repository managers ensure their repositories are harvested?

- Ensure that the metadata is available in a simple Dublin Core datastream. This is the minimal requirement for harvesters.
- Ensure your repository is OAI compliant.
- Notify a service provider (e.g. ARROW Discovery Service) of your repository.
- Discuss with your service provider what your expectations for harvesting include and what they can offer.

The [National Science Digital Library \(NSDL\)](#)⁵⁹ advises that best practice is to register with [the official OAI Registry](#)⁶⁰. (this is not always up to date but it is NLA experience and they generally provide a good service. It can be a pointer to details such as platforms used and policies for deposit).

How to register your repository

- For the ARROW Discovery Service, contact [ARROW Discovery Service](#)⁶¹.
- For OAIster, visit <http://oaister.umdl.umich.edu/o/oaister/dataproviders.html>⁶²
- For Elsevier, visit <http://www.scirus.com>⁶³
- The OAI Registry has an [OAI registration page](#)⁶⁴ for each university to register its repository.

What is the relationship between major harvesters? What agreements exist among them?

- NLA/ARROW DS is liaising with a range of other providers. Other providers also contact the NLA/ARROW DS, namely SCIRUS (Elsevier), NERIUS (European), EdNA Online (Oz), PerX (Scotland).
- Since ARROW DS is liaising with external providers, they are able to broker arrangements on behalf of any university which chooses to participate in their services.
- OAIster harvests the ARROW DS so there is no need to register with both. Some universities (e.g. ANU) are registered with OAIster since they were historically ahead of the rest with their harvesting options. The advantage of registering with the ARROW Discovery Service is that your repository is included with the internationally recognised national provider and will automatically be harvested by OAIster anyway. (there is some complexity in the overlaps that have resulted from these different past arrangements with OAIster so that ARROW finds it must exclude data harvested by OAIster if a repository has arrangements with both.)
- Elsevier (the SCIRUS harvester) also allows for agreements direct with individual universities (e.g. QUT, UQ). They are also negotiating with ANU and NLA. They currently only harvest repositories that are on a DSpace or EPrints platform, and of course, that archive predominantly science based material. They merge repository metadata and indexable full text records into single datastreams. This facilitates simultaneous full text and metadata field searching. But it also means that they cannot harvest image, audio or video files. The richness of the metadata record for these nontext files thus becomes all the more important for discovery.
- OCLC harvests from the ADT program for the ND LTD (Network Digital Library of Thesis and Dissertations) Union Catalog.

What is the relationship of harvesters with Google and Yahoo!?

- NLA/ARROW DS liaises with Google, Google Scholar and Yahoo. Google does not harvest OAI-PMH metadata but NLA is one of the many providers who are still trying to persuade them to do this.
- ARROW Discovery Service and OAIster have agreements with Google for Google to harvest their registered repositories.
- OAIster's sends their metadata to Yahoo! and Google on a monthly basis. How Yahoo! and Google use this metadata is explained at <http://www.oaister.org/sru.html>⁶⁵.

What metadata schema do the different harvesters look for?

- Simple Dublin Core is mandatory for the initial harvest but other schemas are encouraged as well
- OAIster is currently exploring MODS
- OCLC's harvest of thesis material supports an NDLTD thesis schema known as [ETD-MS](#)⁶⁶ (Electronic Thesis and Dissertation - Metadata Set) which is not used by ADT. However ADT materials are still harvested by OCLC because ETD-MS includes the use of Dublin Core elements used by ADT. See the thesis metadata spreadsheet in Appendix B of *Entering Metadata: Guides and Tools for Repositories*.

What software tools are available to test repository for harvestability?

- The [OAI site](#)⁶⁷ has a tool at that can be used to test a repository's OAI compliance. It is called OAI Repository Explorer at <http://re.cs.uct.ac.za>⁶⁸
- Be aware, however, that the majority of IRs that are created today are expected to be OAI-compliant by default. The compliance of IRs will be affected by modifying metadata in a repository unless you adhere to existing standards.

Harvesting 3: Australasian Digital Theses

How to ensure ADT Program theses are harvested by ADT

Libraries will want to migrate theses already deposited with the [Australasian Digital Thesis Program](#)⁶⁹ to their repositories. Note that ADT will still expect to be able to harvest those theses separately from the rest of the repository.

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

There are two steps to ensuring ADT harvesting from the repository: configuring the repository and notifying ADT of your configuration.

There is a difference in the configuring stage between repositories archiving their ADT records in Collections and those who do not.

Configuring the repository - without Collections

ADT only harvests a limited range of Dublin Core elements. For a thesis record to be harvested by ADT, the DC datastream for that thesis record needs to consist of the following:

dc.title

dc.creator

dc.subject

dc.description

dc.date

dc.language

dc.publisher

dc.rights

dc.identifier

dc.type

Any other DC elements in an ADT record (e.g. dc.format) they will be ignored by the ADT harvester.

As usual, all DC fields are repeatable, so multiple dc.subject elements and values can be entered.

Simple Dublin Core is recommended because it meets the minimum standards of all OAI compliant harvesters, and one can be confident that any data expressed in unqualified DC will be harvested by all OAI service providers.

dc.date

ADT only requires the date the thesis was awarded. This date should be the only date (expressed as a 4 digit year) that appears in an ADT record's DC data stream.

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

Qualified Dublin Core?

The ADT forms automatically generate their own DC data in HTML format. Some of these ADT generated DC terms are qualified DC, such as `dc.creator.personalName`. If ADT records have been migrated to the repository and these qualified DC terms appear in the DC data stream of the repository, ADT will harvest these as well as the shorter Simple DC elements listed above.

Sets

Open Archive searching requires that ADT records in a repository be grouped and harvested as a discrete "Set" of items. Technical details for constructing sets are explained in the [Sets Guidelines for Repository Implementers](#) on the [Open Archives Initiative](#)⁷⁰ site.

The Sets Guidelines say of Sets:

Sets provide a method of exposing a partitioning of a repository's contents to harvesters. While this allows for sophisticated harvesting, not all harvesters will exploit this capability. Like non-DC metadata formats, sets are most likely to be useful within specific communities. It is recommended that implementers defer the implementation of sets until there is a particular community-specific situation or deployment scenario that needs sets. Implementation of sets is optional; repository implementers may choose not to implement sets (harvesters may also ignore sets).

ADT is such a specific community that justifies and requires the use of Sets for harvesting.

Sets require a `SetName` and a `SetSpec` in the OAI configuration.

It is recommended that the `dc.type` or `dc.relation` element be used for the `SetName` for ADT harvesting. Enter "Australasian Digital Thesis" as the value for `dc.type` or "Australasian Digital Thesis Program" as the value for `dc.relation`. For example:

```
dc.type Australasian Digital Thesis
```

```
dc.relation Australasian Digital Thesis Program
```

Although it is possible to enter a shorter value, such as ADT only, it is preferable to use terms that are clear to all potential users.

After repository managers decide what metadata properties the Sets are to be based on, technical staff can then make the necessary configurations.

In the repository, if one is not using Collections to store the ADT records, it is simplest for technicians to configure using the `dc.type` or `dc.relation` value for both the `SetName` and `SetSpec` required by the ADT harvester.

For Fedora to be able to configure OAI harvesting, the [OAI provider module](#)⁷¹ needs to be downloaded and installed.

Mapping to Dublin Core

Data entered in MARCXML or some other non-DC format will need to be mapped to a DC data stream. Crosswalks and ADT thesis templates for MARCXML and MODSXML will be

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

prepared soon by RUBRIC.

Configuring the repository - with Collections

The difference in configuring a Collection-based repository for ADT harvesting concerns Sets only. This means that `dc.type` will no longer be a critical element, and the allocation of `SetNames` and `SetSpecs` will also be different. The remainder of the information above is still applicable.

If one's repository uses Collections (e.g. DSpace, or the parent-child collections possible in VITAL), then the name of the Collection storing the ADT records will need to be the `SetSpec` value. In DSpace, this will be the handle identifier (e.g. `hdl_2292_2`) for the Collection; and in VITAL it will be the `dc.identifier` value (e.g. `rubric:299`) that has been assigned for the Collection Parent record.

Informing ADT

For ADT to harvest an OAI-PMH compliant repository, repository managers will need to inform ADT of:

- the URL of your server
- the `SetSpec` of the ADT records to be harvested
- the `SetName` for the ADT records to be harvested (i.e. `dc.type` Australasian Digital Thesis)

Examples:

DSpace record

- URL: <http://researchspace.auckland.ac.nz/dspace-oai/request>
- `SetSpec`: `hdl_2292_2`
- `SetName`: PhD Theses

VITAL record without Collections

- URL: <http://repository.usq.edu.au/oaiprovider>
- `SetSpec`: Australasian Digital Thesis
- `SetName`: Australasian Digital Thesis

VITAL record with ADT items in a Collection

- URL: <http://repository.usq.edu.au/oaiprovider>
- `SetSpec`: `rubric:299`
- `SetName`: Australasian Digital Thesis

ADT contact address mailtoadt-support@unsw.edu.au

Metadata Management and Quality Control

In order to maintain data quality and standards, the repository manager should decide which staff will be responsible for metadata management. The metadata management person or team will be responsible for communicating and ensuring the maintenance of metadata standards. Poor or incomplete data can result in loss of resource discovery potential and in breakdowns of audits for local administrative purposes.

Self-submission and Metadata

In the earliest days of starting a repository, library staff may be responsible for entering most of the data. As workflows and metadata requirements begin to emerge, more of the initial data entry will devolve to academics and faculty research assistants. This data will need to be checked by qualified staff for accuracy and formatting. Entry of correct titles and dates, identification of the issue of the parent publication and other descriptive metadata should not be assumed.

Quality Control

Metadata editors will also be responsible for enhancing the metadata record with additional information such as copyright and access details, enhanced descriptive metadata for discovery, other information that may be needed for internal reporting or monitoring of the collection, and ensuring that a correct version and format of the resource is deposited with the metadata record. Editorial staff may be required to convert the format of the document into a pdf file.

Management will need to decide the editorial steps and levels of editorial privileges in the workflow in order to monitor the quality of the data in the record before it is finally “made live”. Repository software can generally be configured to sequence such workflow steps.

Regular communication channels should be established to review current policies and standards and their application, issues arising and the need for adjustments to workflow and procedures.

Entering Metadata: Guides and Tools for Repositories

Entering Metadata: Guides and Tools for Repositories contains more detailed information, including:

- metadata entry guidelines
- templates for resource types

- file-naming principles
- metadata conversion tools
- crosswalks between schema
- display configuration worksheet
- links to other tools and resources

Further Guides to Metadata

Guides that have most relevance for repositories in higher education institutions and that may be consulted online are:

- [Best Practices for Sharable Metadata](#)⁷²
- [DC Metadata Thesaurus for Repositories](#)⁷³
- [IMS Best Practice Guide](#)⁷⁴

Criteria for evaluating a metadata schema for a digital repository can be found at the following websites:

- [Choosing a Metadata Standard for Research Discovery](#)⁷⁵ (UKOLN 2006) discusses an up to date checklist of guidelines to assist one in choosing a metadata schema for a repository.
- [Metadata Schemes Points of Comparison](#)⁷⁶
- [Repository Librarian and the Next Crusade: The Search for a Common Standard for Digital Repository Metadata](#)⁷⁷ (Goldsmith B B & Knudson, F 2006) - a comparative evaluation of MARCXML, Dublin Core, ONIX, PRISM and MODS.

References and Further Reading

Refer to the Further Reading section at the end of the Toolkit for bibliographic details of works referenced in this section.

“RUBRIC Toolkit: Metadata” produced May 2007

78





Regional Universities Building Research Infrastructure Collaboratively

<http://www.rubric.edu.au/>

Copyright⁷⁹ 2007 RUBRIC⁸⁰

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

- 1 <http://dublincore.org/documents/dces/>
- 2 <http://www.loc.gov/marc/>
- 3 <http://www.agls.gov.au/>
- 4 <http://www.e.govt.nz/standards/nzxls/standard>
- 5 <http://www.gils.net/>
- 6 <http://www.loc.gov/ead/>
- 7 <http://www.vraweb.org/projects/vracore4/index.html>
- 8 <http://www.loc.gov/standards/mix/>
- 9 http://www.getty.edu/research/conducting_research/standards/cdwa/moreinfo.html
- 10 <http://libraries.mit.edu/guides/subjects/metadata/standards/tei.html>
- 11 <http://www.ndltd.org/standards/metadata/current.html>
- 12 http://ethos toolkit.rgu.ac.uk/wp-content/ethos-content/UKETD_DC.htm
- 13 <http://metalogger.files.wordpress.com/2007/06/tcl-etd-mods-profile.pdf>
- 14 http://www.d-nb.de/eng/standards/pdf/ref_xmetadiss_v1-3.pdf
- 15 <http://www.abes.fr/abes/documents/tef/recommandation/index.html>
- 16 <http://www.dublincore.org/documents/2000/07/11/dcmes-qualifiers/>
- 17 <http://www.loc.gov/standards/mods/>
- 18 <http://hul.harvard.edu/jhove/>
- 19 <http://www.loc.gov/standards/premis/information-sheet.pdf>
- 20 <http://www.oclc.org/>
- 21 <http://www.rlg.org/>
- 22 <http://www.openarchives.org/>
- 23 <http://www.oaforum.org/tutorial/english/page6.htm>
- 24 <http://search.arrow.edu.au/>
- 25 <http://www.oaister.org/>
- 26 <http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-63/html/>
- 27 <http://eprints.rclis.org/archive/00005544/01/comparingschemes.pdf>
- 28 <http://dublincore.org/documents/usageguide/index.shtml>
- 29 <http://www.openarchives.org/>
- 30 <http://dublincore.org/index.shtml>
- 31 <http://www.loc.gov/marc/marcdocz.html>
- 32 <http://www.openarchives.org/>
- 33 <http://www.loc.gov/marc/marcdocz.html>
- 34 <http://oai-best.comm.nsd.gov/cgi-bin/wiki.pl?MultipleMetadataFormats>
- 35 <http://acronyms.thefreedictionary.com/AACR2R>
- 36 <http://simile.mit.edu/>
- 37 <http://www.loc.gov/standards/mods/>
- 38 <http://oai-best.comm.nsd.gov/cgi-bin/wiki.pl?MultipleMetadataFormats>
- 39 <http://simile.mit.edu/>
- 40 <http://www.loc.gov/marc/ndmso.html>
- 41 <http://dublincore.org/documents/usageguide/qualifiers.shtml>
- 42 <http://oai-best.comm.nsd.gov/cgi-bin/wiki.pl?MultipleMetadataFormats>

43 <http://dublincore.org/>

44 <http://dublincore.org/documents/library-application-profile/index.shtml>

45 <http://dublincore.org/>

46 <http://dublincore.org/index.shtml>

47 <http://www.pictureaustralia.org/>

48 <http://www.loc.gov/standards/mods/registry.php>

49 <http://metalogger.wordpress.com/2007/08/10/recommended-australian-digital-thesis-metadata/>

50 <http://www.aqf.edu.au/pdf/handbook.pdf>

51 <http://www.aqf.edu.au/>

52 <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/0350530501.html>

53 <http://search.arrow.edu.au/>

54 <http://www.arrow.edu.au/docs/files/harvesting.pdf>

55 <http://dublincore.org/documents/usageguide/elements.shtml>

56 <http://search.arrow.edu.au/>

57 <http://oaister.umdl.umich.edu/o/oaister/>

58 <http://www.nla.gov.au/pub/gateways/archive/76/Campbell-ArrowProject.html>

59 <http://oai-best.comm.nsd.l.org/cgi-bin/wiki.pl?RepositoryLifecycle>

60 <http://www.openarchives.org/Register/BrowseSites>

61 <http://search.arrow.edu.au/>

62 <http://oaister.umdl.umich.edu/o/oaister/dataproviders.html>

63 <http://www.scirus.com/>

64 <http://www.openarchives.org/data/registerasprovider.html>

65 <http://www.oaister.org/sru.html>

66 <http://www.ndltd.org/standards/metadata/current.html>

67 <http://openarchives.org/>

68 <http://re.cs.uct.ac.za/>

69 <http://adt.caul.edu.au/>

70 <http://www.openarchives.org/>

71 <http://www.fedora.info/download/2.1/userdocs/server/features/serviceframework.htm>

72 <http://oai-best.comm.nsd.l.org/cgi-bin/wiki.pl?PublicTOC>

73 https://rubric-central.usq.edu.au/projects/trac/rubric/attachment/wiki/MetadataForHarvesting/rubric_md_thesaurus.pdf?format=raw

74 http://www.imsglobal.org/metadata/mdv1p3pd/imsmd_bestv1p3pd.html

75 <http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-63/html/>

76 <http://eprints.rclis.org/archive/00005544/01/comparingschemes.pdf>

77 <http://www.dlib.org/dlib/september06/goldsmith/09goldsmith.html>

78 <http://creativecommons.org/licenses/by-sa/2.5/au/>

79 <http://creativecommons.org/licenses/by-sa/2.5/au/>

80 <http://www.rubric.edu.au/>