



Picture Australia Contributors Guide to Metadata Harvesting

Version 1.0

Last updated August 26, 2009
(To be revised 2011)

Table of Contents

Picture Australia Contributors Guide to Metadata Harvesting	1
Table of Contents	2
Introduction and background	3
Purpose of this document.....	3
How does Picture Australia work?.....	3
What is Metadata in Picture Australia?	3
Principles of Good Metadata.....	4
Section 1: What do I need to do to my records to get them into Picture Australia?	5
Guidelines for the Picture Australia Metadata Schema	5
Element usage in Picture Australia: Dublin Core element descriptions.....	6
Element usage in Picture Australia: Picture Australia element descriptions	12
Sample of record adhering to the Picture Australia Metadata Schema	13
Converting MARC records to Dublin Core format.....	15
Section 2 : How does Picture Australia get my records?	16
1. Metadata harvesting using OAI-PMH	16
OAI requests.....	17
Other relevant information.....	18
Character encoding: Unicode essential.....	18
2. OAI PMH Static Repository harvesting	19
3. Harvest Control Lists.....	19
4. Other harvesting methods	20
5. The Harvesting Process	20
Contact.....	21
References and further sources of information	22

Introduction and background

Purpose of this document

The purpose of this document is to provide Picture Australia contributors with the information they need to get their records into the Picture Australia service. The document structure is:

1. General introduction to Picture Australia and metadata in Picture Australia.
2. What do I need to do to my records to get them into Picture Australia? This section looks at the Picture Australia Metadata Schema.
3. How does Picture Australia get my records? This section looks at harvesting records through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

How does Picture Australia work?

[Picture Australia](#)¹ is an online service that provides access to distributed image collections containing material of relevance to Australians. Records of images in the Service are provided by [Picture Australia contributors](#)².

Summary of how the service works:

1. Contributors join Picture Australia.
2. Contributors ensure that the metadata for their images complies with the Picture Australia Metadata Schema.
3. Contributors make the metadata about their images available for harvesting via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or Harvest Control Lists (HCL).
4. The National Library harvests metadata using a custom built metadata harvester. The harvested metadata records are then made available through the Picture Australia service. Each record in the Picture Australia service links back to the full record (and image) in the digital repository of the contributing institution.
5. Regular update harvests are conducted to collect new or changed records.

What is Metadata in Picture Australia?

Metadata is structured data about data. It is a summary of information about the form and content of a resource. Descriptive metadata describes a resource for purposes such as discovery and identification similar to the way a catalogue record is used to describe an item on a library shelf. A metadata record consists of a number of pre-defined elements representing values of a resource, and each element can have one or more values. Below is an example of a descriptive metadata record:

Element name	Value
Title	Sydney Harbour Bridge
Creator	Healy, Francis Robert
Description	A view from the Sydney Harbour Bridge

	looking towards Sydney on the right hand side of the bridge, January 1936.
Format	Photograph 6.3 cm x 3.8 cm
Date	1936

In Picture Australia, descriptive metadata assists users of the online service to discover images and photographs for various purposes including study, retrieval, re-use and general interest.

Principles of Good Metadata

Quality metadata is particularly important as it increases the likelihood that digital content will be discovered and used. In [A Framework of Guidance for Building Good Digital Collections](#)³ the National Information Standards Organization (NISO) articulates six principles applying to good metadata:

- Good metadata conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential future uses of the collection.
- Good metadata supports interoperability.
- Good metadata uses authority control and content standards to describe objects and collocate related objects.
- Good metadata includes a clear statement of the conditions and terms of use for the digital object.
- Good metadata supports the long-term curation and preservation of objects in collections.
- Good metadata records are objects themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification.

To assist the creation of quality metadata, the Picture Australia Metadata Schema uses the Dublin Core Metadata Schema and extends it with three elements. The elements used in the schema provide a framework for descriptive metadata.

The [Dublin Core](#)⁴ Schema was designed by the [Dublin Core Metadata Initiative](#)⁵, an open organization that provides simple standards to facilitate the finding, sharing and management of information. The semantics of Dublin Core have been established by an international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields of scholarship and practice.

The ‘elements’ in the Dublin Core schema define what is recorded about each item and in what way. Dublin Core has an unqualified (simple), and a qualified, element set. The unqualified set, created as a means to express a limited set of metadata values pertaining to any digital resource, is used in the Picture Australia Metadata Schema.

Section 1: What do I need to do to my records to get them into Picture Australia?

Guidelines for the Picture Australia Metadata Schema

The Picture Australia metadata schema consists of the fifteen unqualified Dublin Core elements plus three extra Picture Australia elements. To get your records into Picture Australia, the information for each image needs to be encoded in XML that conforms to the Picture Australia metadata schema.

The fifteen unqualified Dublin Core elements are:

- contributor
- coverage
- creator
- date
- description
- format
- **identifier (m)**
- language
- publisher
- relation
- rights
- source
- subject
- **title (m)**
- type

The three additional Picture Australia elements are:

- **pa:thumbnail (m)**
- pa:viewcopy
- **pa:location (m)**

Note:

***(m) denotes mandatory elements**

All Dublin Core elements are repeatable.

Picture Australia elements are not repeatable.

Element usage in Picture Australia: Dublin Core element descriptions

Dublin Core element:	contributor
-----------------------------	--------------------

Element displayed in Picture Australia web: contributor

Definition: An entity responsible for making contributions to the content of the resource. Examples of a contributor include a person, an organization or a service.

Obligation: optional

Guidelines for the creation of element content:

- Is **not necessarily** the contributor of the record.
- Has component parts which should be listed in order of: surname, forename and/or initials, honorific, dates.
- Each separate creator should be listed in a separate occurrence of the element. If several creators are listed in one string, they will appear as a single creator in Picture Australia. The component parts should be separated by commas.
- Each separate contributor should be listed in a separate occurrence of the element. If several contributors are listed in one string, they will appear as a single contributor in Picture Australia.
- Can include dates.
- Can be derived from a Name Authority File.

Correct example:

`<dc:contributor>Laplace, Cyrille, Dr., (1918-1972)</dc:contributor>`

Correct example:

`<dc:contributor>W. A. Jones & Co. </dc:contributor>`

`<dc:contributor>State Library of Victoria</dc:contributor>`

Incorrect example:

`<dc:contributor> W. A. Jones & Co. ; State Library of Victoria</dc:contributor>`

Dublin Core element:	coverage
-----------------------------	-----------------

Element displayed in Picture Australia web: date or place

Definition: Geographical and chronological aspects of the content of the work (essentially, date and place). Coverage will typically include spatial location (a place name or geographic co-ordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of co-ordinates or date ranges.

Obligation: optional

Guidelines for the creation of element content:

- Date format should conform to [ISO 8601](#)⁶.
- Each separate coverage data should be listed in a separate occurrence of the element.
- For circa dates use c. followed by the date e.g. c. 1831
- For uncertain dates use a question mark after the date e.g. 1831?

Correct example (spatial):

`<dc:coverage>Indian Ocean</dc:coverage>`

Correct example (temporal):

`<dc:coverage>1830-1832</dc:coverage>`

Correct example (temporal):

<dc:coverage>c.1830-1832?</dc:coverage>

Incorrect example:

<dc:coverage> Indian Ocean , 1830-1832</dc:coverage>

Dublin Core element:	creator
-----------------------------	----------------

Element displayed in Picture Australia web: creator

Definition: An entity primarily responsible for making the content of the resource. Examples of a Creator include a person, an organisation, or a service.

Obligation: optional

Guidelines for the creation of element content:

- Creators should be listed in the following order: surname, forename and/or initials, honorific, dates. The component parts should be separated by commas.
- A single creator should be listed entirely within a single occurrence of the element. Do not list the surname in one occurrence, and the forename in another, for example.
- Birth and death dates can be included.
- Each separate creator should be listed in a separate occurrence of the element. If several creators are listed in one string, they will appear as a single creator in Picture Australia.
- Can be derived from a Name Authority File.

Correct example:

<dc:creator>Bauer, Ferdinand G, Dr</dc:creator>

Correct example:

<dc:creator>Bauer, Ferdinand G, Dr</dc:creator>

<dc:creator>Bridge, Ellis</dc:creator>

Correct example:

<dc:creator> Bridge, E. (1912-1980)</dc:creator>

Incorrect example:

<dc:creator>Dr Ferdinand G Bauer</dc:creator>

Incorrect example:

<dc:creator>Bauer, F. & Bridge, E.</dc:creator>

Incorrect example:

<dc:creator>Bridge</dc:creator>

<dc:creator>Ellis</dc:creator>

Dublin Core element:	date
-----------------------------	-------------

Element displayed in Picture Australia web: harvested but not displayed

Definition: A date associated with an event in the life cycle of the resource, typically the date of publication and/or creation of the work. Recommended best practice for encoding the date value is defined in a profile of ISO 8601.

Obligation: optional

Guidelines for the creation of element content:

- Date format should conform to [ISO 8601](#)⁷
- Multiple dates should be listed in separate occurrences of the element.

Correct example:

<dc:date>1939-02-17 10:30</dc:date>

Correct example:

<dc:date>1835</dc:date>

Incorrect example:

<dc:date>17-02-1939</dc:date>

Incorrect example:

<dc:date>17/02/1939</dc:date>

Dublin Core element:	description
-----------------------------	--------------------

Element displayed in Picture Australia web: description

Definition: a brief summary of the work, an abstract. Description may include but is not limited to: an abstract, table of contents.

Obligation: optional

Guidelines for the creation of element content:

- This element should not be used for rights information; information about the creator; information about associated resources; or any other information that does not pertain to a description of the resource itself.
- Multiple descriptions should be listed in a separate occurrence of the element.
- May include a date if sourced from a caption.

Correct example:

<dc:description> Engraved in plate l.l.: Ferd. Bauer. Engraved in plate l.r.: Brown prod.fl.nov.holl.p.270.n.6. Title engraved in plate l.l. Plate no. 6 of: Illustrationes florae Novae Hollandiae / Ferdinand Bauer. Engraved in plate l.l.: Ferd. Bauer. Engraved in plate l.c.: Brown prod. fl.nov.holl.p.394.n.17. Title engraved in plate l.c. Plate no. 3 of: Illustrationes florae Novae Hollandiae / Ferdinand Bauer. Plates individually catalogued and digitised.

</dc:description>

Correct example:

<dc:description>Condition: good</dc:description>

Incorrect example:

<dc:description>Some rights reserved</dc:description>

Dublin Core element:	format
-----------------------------	---------------

Element displayed in Picture Australia web: format

Definition: The physical or digital manifestation of the resource. Typically, format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration.

Obligation: optional

Guidelines for the creation of element content:

- Multiple formats should be listed in a separate occurrence of the element.

Correct example:

<dc:format>image</dc:format>

Correct example:

<dc:format> Photograph; 16.5 cm x 21.5 cm</dc:format>

Correct example:

<dc:format>stencil</dc:format>

<dc:format>poster</dc:format>

Dublin Core element:	identifier
-----------------------------	-------------------

Element displayed in Picture Australia web: image number

Definition: an unambiguous reference which allows the work to be found in perpetuity. Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system e.g. Image Number. This can be a URL.

Obligation: MANDATORY

Guidelines for the creation of element content:

- If multiple identifier elements are received, Picture Australia will only display the first identifier element.

Correct example:

`<dc:identifier>30328102131488/1</dc:identifier>`

Correct example:

`<dc:identifier> Accn No: NGA 76.946 NGA IRN: 39406</dc:identifier>`

Correct example:

`<dc:identifier> http://nla.gov.au/nla.pic-an23817143</dc:identifier>`

Dublin Core element:	language
-----------------------------	-----------------

Element displayed in Picture Australia web: harvested but not displayed

Definition: A language of the intellectual content of the resource. Recommended best practice for the values of the Language element is defined by RFC 1766. For example, "en" for English, "fr" French or "en-au" for English used in Australia.

Obligation: optional

Guidelines for the creation of element content:

- If the content is in more than one language, the element may be repeated.

Correct example:

`<dc:language>en-au</dc:language>`

Correct example:

`<dc:language>en</dc:language>`

`<dc:language>fr</dc:language>`

Incorrect example:

`<dc:language>English</dc:language>`

Dublin Core element:	publisher
-----------------------------	------------------

Element displayed in Picture Australia web: publisher

Definition: The entity responsible for making the resource available. Examples of a publisher include a person, an organisation, or a service. Typically, the name of a publisher should be used to indicate the entity.

Obligation: optional

Guidelines for the creation of element content:

- Is **not necessarily** the contributor of the record.
- Has component parts which should be listed in order of: surname, forename and/or initials, honorific, dates.
- Each separate creator should be listed in a separate occurrence of the element. If several creators are listed in one string, they will appear as a single creator in Picture Australia. The component parts should be separated by commas.

- Each separate contributor should be listed in a separate occurrence of the element. If several contributors are listed in one string, they will appear as a single contributor in Picture Australia.
- Can include dates.
- Can be derived from a Name Authority File.

Correct example:

<dc:publisher>National Library of New Zealand / Te Puna Mātauranga o Aotearoa</dc:publisher>

Incorrect example:

<dc:publisher>1901</dc:publisher>

Dublin Core element:	relation
----------------------	----------

Element displayed in Picture Australia web: collection or series

Definition: A reference to a related resource; a version of the work, a significant part of the work, a related work. Qualifiers should be provided where possible e.g. “isPartof”. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

Obligation: optional

Guidelines for the creation of element content:

- Multiple relations should be listed in separate occurrences of the element.
- Qualifiers should be provided where possible, for example ‘isPartof’.
- Can be used for URLs that link to the relation data.

Correct example:

<dc:relation>IspartOf Ruth Hollick collection.</dc:relation>

Correct example:

<dc:relation>
http://www.ngv.vic.gov.au/collection/pub/artistItemListing?artistID=3677
</dc:relation>

Dublin Core element:	rights
----------------------	--------

Element displayed in Picture Australia web: rights

Definition: conditions for use and an expression of moral rights and/or copyright in regard to the resource. Copyright should be asserted if known. Could appear as free text or URI e.g. link to a [Creative Commons](#)⁸ licence.

Guidelines for the creation of element content: Management of the digital collections remain with contributors. If an item (‘work’) is in the [public domain](#)⁹ contributing collecting institutions are strongly encouraged to identify this in the dc:rights element. If the contributing agency is the creator of the work, the selection of a Creative Commons licence as a universally recognised code is encouraged.

Obligation: optional

Correct example (contributor rights):

<dc:rights>Reproduction rights owned by the State Library of Victoria</dc:rights>

Correct example (creative commons rights):

<dc:rights>Copyright owned by Kelvin Rowley. Permission for limited re-use is provided under the terms of the Australian Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd) licence. </dc:rights>

<dc:rights>http://creativecommons.org/licenses/by-nc-nd/2.1/au/ </dc:rights>

Correct example (creator rights):

<dc:rights>© All rights reserved</dc:rights>

Dublin Core element:	source
-----------------------------	---------------

Element displayed in Picture Australia web: managed by

Definition: a reference to a resource from which the present resource is derived.

Obligation: optional

Guidelines for the creation of element content: the Picture Australia contributor providing the resource.

Correct example:

<dc:source>Swinburne University of Technology</dc:source>

Incorrect example:

<dc:source>Item held by Monash University</dc:source>

Dublin Core element:	subject
-----------------------------	----------------

Element displayed in Picture Australia web: subject

Definition: The topic of the content of the resource.

Obligation: optional

Guidelines for the creation of element content: Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme. Thesauri provide lists of preferred terms to use as subject headings to try to provide a standard way of describing (and ultimately searching for) items. For example *bushranger* may be listed as the term preferred to *bush ranger*, *bushrangers*, *highwayman* or *highwaymen*. This helps ensure that everyone who uses the thesaurus, describes and searches for *bushranger* rather than any other term.

Picture Australia recommends the [Australian Pictorial Thesaurus](#)¹⁰ (APT) as the preferred thesaurus for the service. The APT provides contemporary Australian terminology for the description of images and its use will ensure the common description of pictorial collections across Australian archives, libraries, museums and galleries.

Use of the APT is not mandatory. However, agencies that are starting out in the description process are encouraged to consider using the APT and those agencies with alternate existing practices may consider extending their scope to include APT terms.

- Multiple subjects should be listed in separate occurrences of the element.
- Commas or semi-colons should not be used to separate multiple subjects.
- Where thesauri have been used, do not include qualifiers explaining which thesauri have been used.

Correct example:

<dc:subject>Wildflowers</dc:subject>

<dc:subject>Australia</dc:subject>

Incorrect example:

<dc:subject>Wildflowers, Australia</dc:subject>

Incorrect example:

<dc:subject>Waratah;Wildflowers;Australia</dc:subject>

Dublin Core element:	title
-----------------------------	--------------

Element displayed in Picture Australia web: title

Definition: a formal name for the resource

Obligation: MANDATORY

Guidelines for the creation of element content:

- Free text is expected in this element.
- If multiple title elements are received, Picture Australia will only display the first title element.
- Records without a title element will not be included in Picture Australia.

Correct example:

<dc:title>Blinky Bill puppet </dc:title>

Correct example:

<dc:title>Title page of Illustrationes Florae Novae Hollandiae</dc:title>

Dublin Core element:	type
-----------------------------	-------------

Element displayed in Picture Australia web: harvested but not displayed

Definition: how the form of the work is packaged. The nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content.

Obligation: optional

Guidelines for the creation of element content:

- Value of the element is generally ‘image’ for Picture Australia.

Correct example:

<dc:type>image</dc:type>

Element usage in Picture Australia: Picture Australia element descriptions

Picture Australia element:	thumbnail
-----------------------------------	------------------

Element displayed in Picture Australia web: thumbnail image

Definition: the URL of a thumbnail sized version of the image hosted on your server.

Obligation: MANDATORY

Repeatable: no

Guidelines for the creation of element content:

- Thumbnail image should be 150 pixels on the longest edge.
- The URL should be a persistent identifier if possible.

Correct example:

<pa:thumbnail><http://www.statelibrary.vic.gov.au/platebk/0/0/0/tn/pb000060.jpg>

</pa:thumbnail>

Picture Australia element:	viewcopy
-----------------------------------	-----------------

Element displayed in Picture Australia web: used for slideshow trails
Note: previously this element was named ‘mediumresolution’ but this is being phased out in favour of a more descriptive element name. Mediumresolution will continue to be accepted but is being depreciated.
Definition: The URL of a medium sized version of the image hosted on your server
Obligation: optional
Repeatable: no
Guidelines for the creation of element content: Viewcopy images have been introduced for the display of picture trails as projected slideshows. The recommended sizes for viewcopy images approximates 600 - 1000 pixels wide for landscape images and 300 - 700 pixels high for portrait images. The slideshow trail design accommodates the various sizes of images supplied by participating agencies Online Public Access Catalogues (OPAC).
Correct example:
<pa:viewcopy>
<http://www.statelibrary.vic.gov.au/platebk/0/0/0/tn/pb000060.jpg>
</pa:viewcopy>

Picture Australia element:	location
-----------------------------------	-----------------

Element displayed in Picture Australia web: link from thumbnail image to contributor’s record
Definition: The URL of a web page for the image (the record)
Obligation: MANDATORY
Repeatable: no
Guidelines for the creation of element content:
o The URL should be a persistent identifier if possible.
Correct example:
<pa:location><http://www.slv.vic.gov.au/miscpics/0/2/2/doc/mp022287.shtml></pa:location>

Sample of record adhering to the Picture Australia Metadata Schema

Sample record from the National Library of Australia

National Library of Australia element	Dublin Core element	Example metadata elements
Contributor	<dc:contributor>	<dc:contributor>Sainson, Louis Auguste de, 1801-1887</dc:contributor>
Coverage spatial	<dc:coverage>	<dc:coverage>Indian Ocean (ocean)</dc:coverage>
Coverage spatial	<dc:coverage>	<dc:coverage>Tung Hai (sea)</dc:coverage>

Coverage temporal	<dc:coverage>	<dc:coverage>1830-1832</dc:coverage>
Creator	<dc:creator>	<dc:creator>Laplace, Cyrille Pierre Theodore, 1793-1875</dc:creator>
Date	<dc:date>	<dc:date>1835</dc:date>
Description	<dc:description>	<dc:description>Part of Voyage autour du monde par les mers de l'Inde et de la Chine de la corvette de sa Majeste La Favorite execute pendant les annees 1830, 1831, 1832 sous le commandement de M. Laplace ... : album historique</dc:description>
Description	<dc:description>	<dc:description>Condition: good</dc:description>
Description	<dc:description>	<dc:description>The Album historique accompanies : Voyage autour du monde par les mers de l'Inde et de Chine execute pendant les annees 1830, 1831, 1832 sous le commandement de M. Laplace ... /C. Laplace. Paris : Imprimerie Royale, 1833-1835.</dc:description>
Description	<dc:description>	<dc:description>Ferguson 1669</dc:description>
Description	<dc:description>	<dc:description>Also available in an electronic version via the Internet at: http://nla.gov.au/nla.pic-an23739973 </dc:description>
Format	<dc:format>	<dc:format>1 print ; 50 cm.</dc:format>
Identifier	<dc:identifier>	<dc:identifier>nla.pic-an23739973</dc:identifier>
Language	<dc:language>	<dc:language>fr</dc:language>
Publisher	<dc:publisher>	<dc:publisher>Paris : Arthus Bertrand</dc:publisher>
Relation	<dc:relation>	<dc:relation> http://www.nla.gov.au/apps/cdview?pi=nla.pic-vn3808360 </dc:relation>
Rights Management	<dc:rights>	<dc:rights>Public Domain</dc:rights>
Source	<dc:source>	<dc:source>Item held by National Library of Australia</dc:source>
Subject	<dc:subject>	<dc:subject>Laplace, Cyrille Pierre Theodore, 1793-1875 -- Journeys -- Pictorial works.</dc:subject>
Subject	<dc:subject>	<dc:subject>Laplace, Cyrille Pierre Theodore, 1793-1875 -- Journeys -- Pictorial works.</dc:subject>
Subject	<dc:subject>	<dc:subject> Voyages around the world -- Pictorial works.</dc:subject>
Title	<dc:title>	<dc:title>Title page of Voyage autour du monde par les mers de l'Inde et de la Chine de la corvette de sa Majeste La Favorite execute

		pendant les annees 1830, 1831, 1832 sous le commandement de M. Laplace ... Album historique [picture] / grave et publie par les soins et sous la direction de M. de Sainson, dessinateur du voyage de l'Astrolabe</dc:title>
Type	<dc:type>	<dc:type>Image</dc:type>
Thumbnail	<pa:thumbnail>	<pa:thumbnail>http://nla.gov.au/nla.pic-an23739973-t</pa:thumbnail>
Viewcopy	<pa:viewcopy>	<pa:viewcopy>http://nla.gov.au/nla.pic-an23739973-v</pa:viewcopy>
Location	<pa:location>	<pa:location>http://nla.gov.au/nla.pic-an23739973</pa:location>

Converting MARC records to Dublin Core format

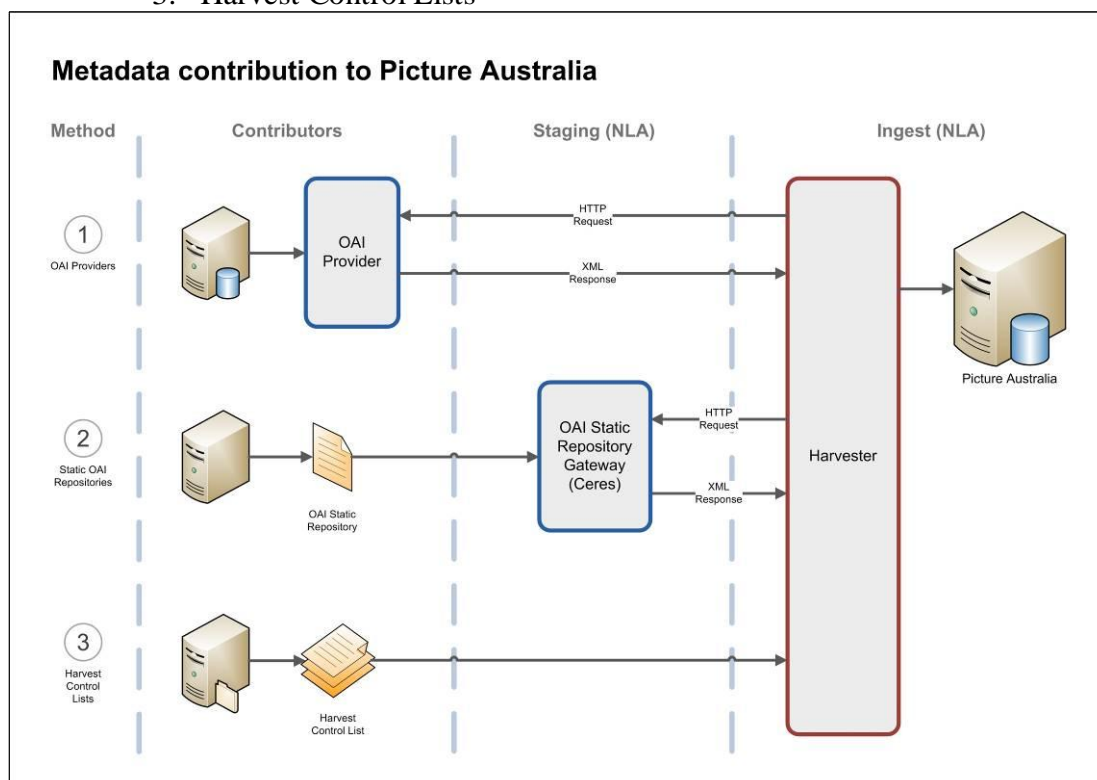
Read this section if you require information on converting MARC records to Dublin Core format.

Contributors joining Picture Australia either create new metadata for their image collections or convert their existing descriptive data into the Dublin Core format. This requires the existing information fields to be mapped to the most appropriate DC elements, as part of a conversion process. Please note that the National Library of Australia does not undertake this process on behalf of contributors. Guidelines for converting MARC records to Dublin Core are maintained by the Library of Congress in the [MARC to Dublin Core Crosswalk](#)¹¹.

Section 2 : How does Picture Australia get my records?

Picture Australia uses any of three methods to harvest and ingest records as shown in the diagram below:

1. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
2. OAI Static Repository
3. Harvest Control Lists



1. Metadata harvesting using OAI-PMH

The metadata for the Picture Australia service is harvested from participating agencies' web sites using the [Open Archives Initiative – Protocol for Metadata Harvesting](#)¹². OAI-PMH is a web-based protocol established in 2001 to define a standard way to move metadata from point A to point B via the Internet. The OAI provides rules and a framework for sharing descriptive metadata, both for making metadata available and for acquiring metadata records once they are made available.

Data can be made available in different metadata formats, but to comply with the protocol, at least unqualified Dublin Core must be supported. Requests for records sent via OAI-PMH specify the metadata schema that is required, and if that schema is supported by the responding repository, records will be returned in that schema.

In the case of Picture Australia a metadata format for the Picture Australia Schema needs to be created or configured in a contributors repository.

The NLA uses a custom-built harvester application (released as open source and available from <http://code.nla.gov.au>¹³) to issue OAI requests and collect metadata from various repositories and other collections of data. This metadata is then made available in one of several ways:

- by a user-driven web interface;
- via the Open Search protocol;
- via the z39.50 protocol; and
- via OAI-PMH used by other data aggregators.

How harvesting works:

1. the NLA harvester issues an OAI request via http to an OAI compliant repository.
2. the request is responded to by the repository with metadata encoded in [XML](#)¹⁴ in the appropriate metadata format.

OAI requests

The OAI-PMH defines six requests that are used by metadata harvesters to request information from repositories. Each request is distinctive in purpose and meaning. Here is a brief description of each:

OAI request	OAI response
Identify	Provides basic information about the repository such as repository name, base URL, protocol version, earliest date stamp, granularity, support for deleted records, repository contact e-mail.
ListSets	Provides a list of sets that are established in the repository.
ListMetadataFormats	Provides a list of metadata formats that are supported, for example unqualified Dublin Core.
GetRecord	Provides an individual record in the repository.
ListRecords	Provides the metadata for each record that meets the specified criteria (such as a specific date range).
ListIdentifiers	Provides basic information about each record in the repository that meets the specified criteria, including the Unique Identifier.

For further information on OAI requests and responses see <http://www.openarchives.org/OAI/openarchivesprotocol.html>

The final three requests in the table above can be qualified with filters for searching records:

Date ranges: A request can specify a date range for the returned records (e.g. only records added/deleted or changed after 9am March 17, 2005). Like most harvesters, the National Library's harvester will specify that it only wants details that have changed since the date of the last harvest. When this is done, a small overlap of time is allowed.

Metadata prefix: All GetRecord, ListIdentifiers and ListRecords requests must have a metadata prefix specified.

Set: A request can specify that it only wants records from a specific set. A [set](#)¹⁵ is an optional construct for repositories to group items for the purpose of selective harvesting. Picture Australia does not require any specific sets to be established to enable harvesting. However, if your repository contains records that you do not want to appear in Picture Australia, establishing a set is probably the best way to enable the National Library to harvest only what you want to expose. The procedures for establishing sets vary according to the software application that is being used. The technical support person for your repository should be able to clarify how to do it.

Other relevant information

Unique identifier

OAI records have unique identifiers which are allocated automatically by the repository software. These identifiers are NOT the same as the URL identifiers that Picture Australia uses to link to records in your repository. OAI identifiers occur in the headers of records, and are used by harvesters to update records and disambiguate similar records.

Creation, modification and deletion

To comply with OAI-PMH, your repository will export details of records that have been added (created) and those that have been modified in the time ranges that the harvester request has specified. Some repositories are also enabled to export details of records that have been deleted (while creation and modification must be supported under OAI-PMH, support for deletion information is optional). The National Library recommends that repositories do enable deletion information if possible, as it ensures that Picture Australia does not continue to contain records that no longer exist, which is frustrating for users of the service.

Character encoding: Unicode essential

Because OAI-PMH uses XML encoding to transmit the metadata, it is essential that a repository can export in valid XML, including using UTF-8 character encoding only. UTF-8 is one of the three forms of the [Unicode](#)¹⁶ standard. Data created in programs such as Microsoft Word often contain special characters. If these are not picked up when copied into a repository, then when the data is exported, it will not be recognised as valid XML and the harvest may fail. Use of an XML validator is recommended as it allows you to check your XML documents on conformance to [W3C specifications for XML](#)¹⁷.

2. OAI PMH Static Repository harvesting

A [Static Repository](#)¹⁸ provides a simple approach for exposing relatively static and small collections of metadata records through the OAI-PMH.

A Static Repository is an XML file, available at a persistent HTTP URL that conforms to the rules specified by OAI-PMH. A Static Repository contains metadata records and supporting information required for the purpose of harvesting via the OAI-PMH through the intermediation of a Static Repository Gateway. A Static Repository is not a OAI-PMH Repository, because it is a file, not a server that can respond to the six OAI-PMH requests.

The Static Repository approach is targeted at organisations that:

- Have metadata collections ranging in size between 1 and 5000 records depending on the file size of each record. The National Library limits file size of the Static Repository to 10MB;
- Can make static content available through a network-accessible Web server;
- Need a technically simpler implementation strategy than providing an OAI-PMH repository.

A Static Repository becomes harvestable via the OAI-PMH through the intermediation of a Static Repository Gateway. The National Library has implemented a Static Repository Gateway, known as Ceres, that allows harvesting of static repositories into Picture Australia.

For the purposes of Picture Australia, the data obtained from a static repository will be subject to the same rules and data transformations as data obtained from an OAI-PMH repository.

Note: there are a number of restrictions on a Static Repository relative to a standard OAI-PMH Repository. Static Repositories do not support sets, deleted records, response compression, harvesting granularity other than YYYY-MM-DD, or resumptionTokens.

Full guidelines for static repositories can be found on the OAI website at:

<http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm>

3. Harvest Control Lists

Some contributors to Picture Australia may opt to contribute metadata via a Harvest Control List (HCL). Harvest Control Lists were used before the development of OAI-PMH and this option is not recommended by the National Library because of its limited functional flexibility and negative impact on data currency.

A Harvest Control List is a simple list of links to metadata records conforming to the Picture Australia Metadata Schema.

- An HCL must be published to the web;

- The HCL file should be stored in a permanent location and point to Picture Australia records in XML format conforming to the Picture Australia Schema.

4. Other harvesting methods

If you are unable to provide your metadata records using the above methods, please contact us to discuss your requirements.

5. The Harvesting Process

When your records adhere to the Picture Australia Metadata Schema and are ready for harvesting, National Library staff will employ the test harvester application to conduct a test harvest. The test harvest is run to ensure OAI requests issued from the harvester are responded to correctly and to view the metadata in a secure test environment. After a successful test harvest has been completed and any problems resolved, a production harvest will follow and your records will display in the Picture Australia service at <http://www.pictureaustralia.org>

The harvesting process is as follows:

	Task	Prerequisite	Role
1	Provide OAI base URL/HCL	OAI repository interface or HCL in place	Contributor
2	Harvest contributor's metadata into the Picture Australia test system	Requires written approval from contributor to harvest into test.	NLA
3	Review metadata records and make any necessary changes		NLA/Contributor
4	Repeat steps 2 and 3 until contributor is happy		NLA/Contributor

5	Harvest contributor's metadata into the Picture Australia production system	Requires signed Picture Australia Memorandum of Understanding (MOU)	NLA
6	Review metadata records		NLA/Contributor
7	Repeat steps 2 to 6 if necessary (if there are issues with the review in Prod)		NLA/Contributor
8	Schedule incremental harvests		NLA/Contributor

Contact

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

Email: pictaust@nla.gov.au

References and further sources of information

¹ <http://www.pictureaustralia.org>

² <http://www.pictureaustralia.org/contribute/participants/index.html>

³ <http://framework.niso.org/>

⁴ <http://dublincore.org/documents/dces/>

⁵ <http://dublincore.org/about/>

⁶

http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm

⁷

http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm

⁸ <http://creativecommons.org/>

⁹ <http://creativecommons.org/licenses/publicdomain/>

¹⁰ <http://www.picturethesaurus.gov.au/>

¹¹ <http://www.loc.gov/marc/marc2dc.html>

¹² <http://www.openarchives.org/pmh/>

¹³ <http://code.nla.gov.au>

¹⁴ <http://www.w3.org/XML/>

¹⁵ <http://www.openarchives.org/OAI/openarchivesprotocol.html#Set>

¹⁶ <http://www.unicode.org/standard/WhatIsUnicode.html>

¹⁷ <http://www.w3.org/XML/>

¹⁸ <http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm>