ANDS Guide

Metadata

Level: Working

Last updated: December 2016

Web link: www.ands.org.au/guides/metadata-working

This Guide is intended to provide a simple generic working-level view of the needs, issues, and processes around metadata collection and creation as it relates to research data.

For those needing information and advice about the RIF-CS metadata schema used for Research Data Australia, please see the Content Providers Guide.

Table of Contents

Metadata 1

Key points 2

What is metadata? 2

Metadata for research data 3

Why create metadata? 3

- **Finding data** 3
- **Determining the value of data** 3
- **Accessing data** 4
- **Using and reusing data** 4

Levels of metadata 4

Types of metadata 5

The lingua franca of metadata 6

Elements and schemas **6**
Standards **7**
Content rules and controlled vocabularies **7**
Namespaces **7**
Cross-walking different metadata schemas **7**
Developing your metadata schema **8**
Plan from the user's perspective **8**
Finding and choosing a metadata schema **8**
Adopt, adapt, or create your schema? **8**
Collecting and linking Metadata **9**
Collecting metadata **9**
Linking metadata and data **10**
Metadata within the data file 10
Metadata as a separate file(s) 10
Spreadsheets and databases 10
Specialised metadata and data stores 11
Further reading **11**
Feedback? **11**
About ANDS **11**

Key points

- Richly described metadata is the key to making research data publishable, discoverable, citable and reusable
- Collection, updating and maintaining metadata are necessary inclusions in the planning and budgeting of all research projects
- Because digital data objects often change location, managing the link between data and metadata is critical and there are technologies e.g. persistent identifiers which support this persistent linking of data and metadata
- Use one or several of the plethora of established metadata standards as much as possible: If your project is the only one to use a particular metadata element set, it creates a barrier to interoperability and reuse.

What is metadata?

While generally 'meta-data' is summarised as 'data about data', what does that actually mean?

- Metadata is information about an object or resource that describes characteristics of that object, such as content, quality, format, location and access rights
- Metadata can be used to describe physical objects (pot shards, specimens etc.) as well as digital objects (documents, images, datasets, software etc.)
- Metadata can take many different forms, from free text (e.g. a read-me file) to standardized, structured, machine-readable, extensible content

- Metadata is analogous to any other form of data, in terms of how it is created, managed, linked and stored
- Metadata is associated with the data it describes. It can be embedded within the data file, or recorded a separate text/spreadsheet file that is linked to the collection of data files it describes, or contained in a catalogue record that points to the research data collection.

Metadata for research data

Well described metadata records show the power of rich metadata in making research data collections discoverable, citable, reusable and accessible for the long term.

> Two-Rocks moorings data 2004 - 2005 metadata record in the CSIRO Data Access Portal contains 35 metadata fields which enable researchers to quickly and accurately assess the relevance of this dataset to their research. The metadata record and the data are closely linked through co-location on the same access page. The Files tab contains additional metadata about each of the 17 files within this collection: file type, last modified, and file size.

Rich metadata allows records to be syndicated to other data catalogues; here is the same Two-Rocks mooring data record syndicated to:

- Research Data Australia: Australia's aggregated research data catalogue
- Marlin Oceans and Atmosphere: a discipline-specific metadata catalogue

Why create metadata?

Metadata enables and enhances the discovery and reuse of data.

Finding data

Data formats such as text can be indexed and searched themselves (as in a simple Google search). However, the ability to search formats like audio, images and video is limited, and discoverability relies on searching the metadata. Discovery metadata helps researchers find data that, for example:

- relates to a geographical area of interest (via geospatial metadata)
- relates to a research discipline of interest (via field of research, keyword or vocabulary metadata)
- is generated by another researcher whose work is of interest (via lead researcher or contributor metadata).

Determining the value of data

To assess the usefulness, value and quality of a dataset, researchers need to understand the context around the data. This is given in metadata that:

- Describe why the data were collected, the experimental design and data collection methods, etc.
- Links to the researchers and institution(s) involved
- Identifies the research program or grant
- Points to publications that have flowed from the research data
- Explicitly provides provenance, licensing, rights, and technical information.

Accessing data

Access to research data requires:

- information that identifies the research data collection - usually through a metadata collection (catalogue) record
- Links to the data or contact information:
    - a direct download link to online data for open access, or
    - contact information for the data manager for mediated access.

Using and reusing data

To make use of any dataset, researchers need metadata on:

- how the data is structured
- what it describes
- how to read it (e.g. column headings and units)
- methodological information such as instrument settings and calibrations, reagents used, or survey questions
- exactly what they are allowed to do with the data through rights metadata such as licensing
- how to acknowledge the original creators by citing the data.

Proper recording of this information is important for the producers of the data - it is easy to forget details of experiments that become important when analysing the data - as well as future reusers.

Astronomers are extremely careful in calibrating optical images under different filters that allow only certain colours through. If information about the filters used is separated from the image, the data loses a lot of its value. Similarly, if a photo of a crime-scene has any modifications made, and those changes are not properly recorded, the data loses its credibility and value.

Levels of metadata

There are three levels of data groupings: data objects (**Items and Sub-items**) can be collected into groups (**Collections)**

1. **Sub-item level**. Sometimes single objects can be made up of component parts: a book has chapters, a documentary has scenes, a relational database has tables, and so on. To search for a particular sub-item component requires not just the item-level metadata but also its sub-item or component level metadata.
2. **Item level** metadata describes individual objects
3. **Collection level** metadata describes the collection as a whole. For example, Australian Urban Water Collection

The collection approach and subsequent level of description impacts on discoverability, and cost and effort of management. Metadata ideally should be based on the needs of those for whom the collection is created.

More information: ANDS Guide to *Defining a collection*

EXAMPLE: Collection, item and sub-item metadata:

Every major TV station maintains a video archive of the materials that they have broadcast (such as news, sports, cartoons and documentaries). To support effective use of the archive, the recordings are usually described at several levels:

- *Recordings are grouped into collections based on the type of program – news, sports, soap-operas and so on, rather than say the date of broadcast which may happen more than once, and doesn't tell you anything about the show.*
- *At the item-level they identify the particular episode of a program. For ongoing series they might record metadata elements such as 'Series 10, Episode 5' of 'The Simpsons', which also has a title of 'When you dish upon a star'. For live shows such as News it may be recorded as '7pm News' with the location 'Sydney' and date '1 January 1962'.*
- *At the sub-item level of a News episode they will identify a particular story, e.g. the '7pm News, Sydney, 1 January 1962' item will have a reference to a story on 'Independence for Western Samoa' as being the second story, 5m:14s from the start.*

This hierarchical structure allows users to find particular materials directly, rather than scanning manually through a large number of complete recordings within even a single archive.

Types of metadata

Metadata types are often grouped into functional types, but note that some elements will provide multiple functions.

The most common types are:

- **Descriptive metadata**: information required for discovery and assessment of the collection,
    - e.g. title, contributors, subject or keywords, study description, and location and dates of the study.
- **Provenance metadata**: this relates to the origins and processing of the data, and enables interpretation and reuse of the data. It ranges from the human to the highly technical, and usually requires some knowledge of the domain to create.
    - e.g. Where did the data come from? Why was it collected? Who collected it, when and where? What instruments/technologies were used to collect the data, and how were they set up? How has the data been processed?
- **Technical metadata**: fundamental information for a person or a computer application to read the data.
    - e.g. How is the data set up? What formats, and versions of formats, are used? How is the database configured? How does it relate to other data?
- **Rights and access metadata**: information to enable access, and licensing or usage rules.
    - e.g. How can someone access the data? Who is allowed to view or modify the data, or the metadata, and under what conditions? Who has some kind of authority over the data? Are there costs associated with access? Under what licence is the data being made available?
- **Preservation metadata**: this builds on the history from the Provenance, Rights and Technical metadata, and also includes information to allow the data to be managed for long-term accessibility.

- e.g. Has there been any restructuring or other changes to the files, e.g. due to migration to new file formats? What software has been used to access the data?

- **Citation metadata**: information required for someone to cite the data
    - e.g. Creator(s), Publication Year, Title, Publisher, Identifier.

The lingua franca of metadata

Understanding commonly used metadata terminology will help you better plan, collect and apply metadata. For example: metadata elements, schemas, standards, vocabularies etc.

Elements and schemas

Metadata **schemas** are an overall structure for metadata about a particular information resource or for a specific domain. A schema specifies a set of metadata concepts or terms (called **elements**), and their associated definitions (semantics) and relationships. The value given to each element is the content. Metadata schemas often emerge from a single community group (e.g. the library community) or can be developed to describe a specific format (a digital audio file for example) or domain. For example, MARC is a standard for the representation and communication of bibliographic information in machine-readable form; and ISO 19115 is an international standard for describing geographic information and services.

A schema may also specify:

- **content rules** e.g. required formats and controlled vocabularies
- the **syntax** in which elements must be encoded (or expressed), such as XML (Extensible Markup Language).

However, these are often given in **application profiles**, which are specific implementation rules, and often contain elements from more than one schema. Application profiles may also specify whether an element is mandatory, optional or repeatable.
An example is the Marine Community Profile, which was developed in accordance with ISO 19115 rules by the Australian Ocean Data Centre to support the documentation and discovery of marine spatial datasets.

EXAMPLE

One of the most common generic schemas is the Dublin Core Metadata Initiative, which is the most widely adopted schema for descriptive metadata to date.

It is simple and generic, with just 15 elements in the original Dublin Core Metadata Element Set, including Title, Date, Type, Format, Creator, Coverage and Rights. Dublin Core has since been extended to 55 terms in the DCMI terms namespace.

Each field can be basically free form text, with some restrictions.

Dublin Core is used a lot on web pages, with the "dc" namespace – so you will often see things like "dc.Title" as a metadata tag inside a webpage's html headers. If you are working with video you may use Dublin Core and the MPe.g.-7 standard for video archives and so get the "dc.Title" and the "MPe.g.7.Title" fields. Each has their own rules how you can use them, found in the schema's namespace.

Standards

- **Official**: Metadata schemas that go through a formal validation process by a standards organisation, such as the International Standards Organisation (ISO) or an equivalent body such as the Dublin Core Metadata Initiative (DCMI), become official metadata standards. Examples include the Data Documentation Initiative (DDI), and the PREMIS Data Dictionary for Preservation Metadata.
- '**de-facto**' standards are more common than any official standard and are no less applicable in most cases. Commonly used and consistently applied metadata schemas that are well documented, endorsed, and maintained by someone are also appropriate, e.g. RIF-CS for describing data collections and services used in Research Data Australia.

Content rules and controlled vocabularies

Specifying rules around the allowable content and format of values in each metadata element improves accuracy and machine-readability of metadata, and hence discoverability of collections. Free-form text entry can lead to ambiguous data, for example the date 3/10/15 could refer to 3 October or 10 March, in either 1915 or 2015.

Having a specific set of terms that can be used in a field, i.e. a controlled vocabulary, allows filtering and faceting of the data, improving search function.

Controlled vocabularies can be:

- locally defined (e.g. only allowing names to be selected from a list of employees), or
- an established standard (e.g. the Field of Research Codes for Australian Research subjects, or the Gazetteer of Australia Place Name database).

Likewise, formats can be:

- specified locally (i.e. names will be entered Firstname Lastname), or
- use an international standard (i.e. recommended best practice for the Dublin Core date element is to use an encoding scheme, such as the W3CDTF profile of ISO 8601).

Namespaces

A metadata schema's 'namespace' declares a unique set of elements and definitions. A namespace is ideally expressed as a domain name associated with the schema which, along with the individual element name, produces a URI that uniquely identifies the element. For example, the DCMI namespace is http://purl.org/dc/terms/, and the DC term 'title' is identified and defined by http://purl.org/dc/terms/title.

By specifying the namespace(s) of the metadata schema(s) you are using, you can define which schema each element belongs to, and point people to the accepted definition of that element.

For example, the term "date" appears in many schemas. However, the way a "date" is defined or recorded may be different depending on the schema; "date" might refer variously to the date the data were published in one schema, or the date the data were collected in another schema.

Cross-walking different metadata schemas

A crosswalk maps the elements in one schema to the equivalent elements in another schema in a machine-readable way. This allows a system to ingest metadata that is in a different schema from the one it uses; for

example, the ANDS Registry can harvest metadata encoded in XML from a repository that uses the ISO19115 schema, and using an XSLT (Extensible Stylesheet Language Transformation) crosswalk, transform it into XML-encoded RIF-CS format.

Crosswalking works best if the source metadata content is well structured and formatted consistently.

Developing your metadata schema

Plan from the user's perspective

The metadata that needs to be recorded is based on the needs of users to find, select, access and use the data; as well as the people who will manage and preserve the data. Then you can select which schemas, and associated vocabularies and formats, are the most appropriate.

Finding and choosing a metadata schema

Schemas range from the very generic to extremely discipline or resource specific:

- Generic schemas such as the DCMI are widely adopted and easy to use - but it is so generic that everyone can use it in quite different ways as the description becomes more specialised.
- Discipline-specific schemas provide a richer and more-targeted structure and vocabulary, which allows detailed information to be provided in a more structured, granular format. Finding these can be as simple as searching the internet for "[discipline] metadata schema", from a very high level to start with (e.g. "agriculture metadata schema") and then getting steadily more specific (e.g. "crop yield metadata schema") if required. You can also search metadata schema registries such as the Data Curation Centre's List of Metadata Standards; or the Research Data Alliance's Metadata Standards Directory.

For certain applications, there are discipline-standards bodies, industry groups, or other higher agencies like government departments and UN agencies, which recommend the use of particular schemas and standards:

- Astronomy has the International Virtual Observatory Alliance
- Biodiversity community has Biodiversity Information Standards (TDWG) that is affiliated with the International Union of Biological Sciences
- Agriculture has the Food and Agriculture Organization of the United Nations (FAO)
- Geospatial data metadata standards are being set by the governments of Australia and New Zealand through ANZLIC (The Spatial Information Council)

Adopt, adapt, or create your schema?

Although it is possible to develop a metadata schema from scratch, it is preferable to use or adapt existing standards and/or widely-established schemas, as they offer:

- Cost savings – the schema and its usage guidelines have been developed thus saving time and effort
- Access to help and advice – a standard is likely to have a community of users
- Usability – users are likely to be familiar with a standard and its terminology
- Interoperability - information can be easily shared between systems
- Sustainability – schemas need maintenance and updating if they are to remain usable.

Your choices are:

1) If there is just one obvious metadata standard, and it meets your needs, use it.
2) If there are several obvious schemas that meet your needs, follow models of 'good practice' within your community; keep in mind that well-structured metadata can be mapped or cross-walked from one standard to another.
3) Where you can find no single appropriate schema:
    a) adapt or extend an existing schema to better fit your needs, and document the changes you make very carefully using the documentation methods and mappings deployed by existing standards as a guide. Contact the 'owners' of the schema and attempt to work with them, as others may benefit from your changes.
    b) alternatively, develop a new 'application profile', where various metadata elements (and the elements' guidelines and documentation) are taken from different metadata schemas and mixed together.
4) If there is absolutely no schema you can use, check again; it's a rare situation nowadays. If there is still no schema to be found, then you may have to develop one. However, it takes a fair bit of work and you should bring together as many interested people in your discipline as you can.

Collecting and linking Metadata

Collecting metadata

Automatic metadata collection: A lot of metadata can be created automatically during the data collection process. Many scientific instruments generate metadata alongside the data itself. An obvious example is digital cameras, where some provenance metadata is written at the same time as the photo is taken, e.g. location, time and date. Bigger instruments such as microscopes, particle accelerators and telescopes generate extremely rich and complex metadata during an observation.

Automatic metadata collection avoids data entry errors and reduces the effort required - relying on humans to create metadata can lead to none being created. You can also set up an automated process to sanity-check the metadata when the data comes in.

Extracted metadata collection: Some metadata can be extracted from other systems. For example, a university's human resource management, grant management or research management systems may be the best sources of information regarding researchers, research grants, or research projects.

Human metadata collection: Some metadata requires human participation to create, and the process used depends a lot on the tools used to collect the data, for example metadata captured in electronic notebooks or digital lab books. It is also possible to build tools such as forms, which can present the metadata fields (and even pre-populate some if connected to other institutional systems) and automatically enforce the right vocabularies and data structures for each metadata record. Metadata records can be created as the data is being collected, which is preferable, or can be done later on when the researcher organises the data they have collected over some period of time, e.g. after a field trip.

Metadata is often hidden away in specification statements, database structures, data models, program code or master data reference structures. This metadata needs to be made explicit and human-readable to be useful for humans.

Linking metadata and data

There are a range of options for storing metadata, and not all metadata may end up in the one place. However, searching and managing the metadata is a lot easier if you take a structured approach.

For an institutional view of metadata stores see ANDS Guide: Metadata Stores Solutions

Metadata within the data file

The first place that metadata can go is inside the file with the data itself. There are many digital file formats that include a range of metadata fields, and some can be extended to hold almost anything. These include:

- text formats e.g. DOCX, HTML, and PDF
- image formats e.g. TIFF and JPEG
- video formats e.g. MPEG
- audio formats e.g. WAV
- specialist discipline formats e.g. FITS (Astronomy) and HDF (Remote sensing).

A benefit of storing metadata inside the file is that it moves with the file; the *association* between the data and its metadata is easy to maintain.

However, the downsides are:

- Not every metadata field you want may be able to be added.
- Searching the collection is slow, as the computer has to open every single file for every single query, especially as the number of files and queries grows.
- Collection-level metadata is not easily managed. If you write a collection reference into each file and then decide to make a change to it, you have to edit every single affected file.

Metadata as a separate file(s)

This solution will give you infinite flexibility for storing any and all kinds of metadata without restriction: write the metadata into a separate, well-structured file, (perhaps using XML), and associate that with the data file. A common approach to strengthen the file-metadata file association is to use the same filename stem, e.g. cat1.tiff is the image and cat1.xml is the metadata. This can improve the performance for searching and metadata modifications slightly.

However, the downsides are:

- the data is still on the same storage medium and you still have to open the same number of files to make queries or changes
- it increases the risk when files are moved that the data and the metadata will get separated.

Spreadsheets and databases

Aggregating metadata for multiple datasets into a single spreadsheet or database gives you a lot more flexibility in searching, making changes, and in the metadata fields recorded.

Collection-level changes are very easy to make in a database, as items that belong to a certain collection are just flagged, and pointed to the collection-level record. Only one record (the collection-level record) in the database has to be changed for every item in that collection to be changed.

The metadata is associated with the data by recording the filename and location (or URI) for the data within the metadata record, and updating this anytime the data location is changed.

Specialised metadata and data stores

Data stores or repositories combine a specialised system for metadata capture and storage with file storage systems; while metadata stores hold only metadata, and point to the location of data objects. The benefit of data and metadata stores is that they are usually pre-configured with one or more metadata schemas, and are searchable and accessible with some level of user-friendliness and customisation options. Many can also handle complex metadata, such as non-text formats like AV files or images.

There many options available ranging from free or open source systems, up to large commercial systems. See the ANDS Guide to metadata stores solutions for more information.

Further reading

- Metadata Best Practices. DataONE. http://www.dataone.org/best-practices/metadata
- Metadata Guide, JISC. https://www.jisc.ac.uk/guides/metadata
- Metadata and describing data, Cornell University http://data.research.cornell.edu/content/writing-metadata

Feedback?

We welcome your feedback on this guide. Please email contact@ands.org.au with any comments or questions.

About ANDS

The Australian National Data Service (ANDS) makes Australia's research data assets more valuable for researchers, research institutions and the nation.

ANDS is a partnership led by Monash University in collaboration with the Australian National University (ANU) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). It is funded by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative