**Module 3: Contextual Details Needed to Make Data Meaningful to Others**

**Part I Authors are from Marine Biological Laboratory and Woods Hole Oceanographic Institute (MBLWHOI):**

Elizabeth Coburn

John Furfey

Jen Walton

**Part II Authors are from Tisch Library at Tufts University:**

Alexander May

Alicia Morris

Learner Objectives:

1. Understand what metadata is

2. Understand why metadata is important

3. Identify applicable standards for documenting and capturing metadata

4. Understand disciplinary practices associated with the collection and documentation of metadata

5. Identify an approach to creating metadata for a project

**Part I**

**1. General definitions of metadata**

*Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information.* (2004, NISO, Understanding Metadata pg. 1)

Metadata is used to record information about data (e.g. bibliographic or scientific) that has been collected. Metadata is essential to enabling the use and reuse of data and in ensuring that resources are accessible, and usable, in the future. The Marine Metadata Interoperability Project (MMI) states in its *Introduction to Metadata* (see <https://marinemetadata.org/guides/mdataintro>):

*In today’s research environment, creation of metadata is becoming a requirement for practical use of research observations and results.*

You must have metadata in order to:

* find data from other researchers to support your research;
* use the data that you do find;
* help other professionals to find and use data from your research; and
* use your own data in the future when you may have forgotten details of the research.

Metadata is typically manually created. Some metadata may be collected automatically by scientific instrumentation, as it collects the data.

Metadata is commonly broken down into three main types: descriptive, structural, and administrative.

* Descriptive metadata describes the object or data and gives the basic facts: who created it (i.e. authorship), title, keywords, and abstract.
* Structural metadata describes the structure of an object including its components and how they are related. It also describes the format, process, and inter-relatedness of objects. It can be used to facilitate navigation, or define the format or sequence of complex objects.
* Administrative metadata includes information about the management of the object and may include information about: preservation and rights management, creation date, copyright permissions, required software, provenance (history), and file integrity checks.

There are many different metadata standards and specifications, some of them are discipline, or domain, specific. These standards should be followed to facilitate the successful and continued access to and reuse of data.

The Marine Metadata Interoperability (MMI) Project site has some great examples of metadata records that you should review to gain a better understanding of the concept of metadata (see <https://marinemetadata.org/guides/mdataintro/mdataexamples>).

**Resources:**

1. National Information Standards Organization (NISO). 2004. Understanding Metadata. [http://www.niso.org/publications/press/UnderstandingMetadata.pdf -](http://www.niso.org/publications/press/UnderstandingMetadata.pdf#http://www.niso.org/publications/press/UnderstandingMetadata.pdf)
2. Neiswender, C. 2010. "Introduction to Metadata." In *The MMI Guides: Navigating the World of Marine Metadata*. <http://marinemetadata.org/guides/mdataintro>. Accessed April 1, 2013.

**How metadata facilitates discoverability and reuse**

As has been previously discussed, metadata facilitates discoverability, accessibility, ownership, reuse and data structure by providing necessary information about an object. This information is attached to the object, and will follow it throughout its lifecycle, and facilitate its use. Depending on which metadata scheme is used, and how much about an object is known, the amount of metadata for any object will vary.

Ownership of an object may be indicated in a variety of ways, but typically a user can look at the *creator*, *author*, *publisher*, and *source* elements, or fields of an object’s metadata record. Information about how an object may be reused will typically be indicated by the *rights* element, or field. It may consist of a broad copyright statement, where the owner of the object has decided to retain all rights, or it may consist of licensing information (e.g. a Creative Commons license), which might, for instance, require attribution in exchange for the use of the object. Information about the data structure of an object, if provided, may be indicated by information in the *description* field. For more complicated digital objects (e.g. data sets consisting of more than one file), this may include information about the other files, or the file structure, etc., with more complicated metadata schemas with many more elements (for instance the Metadata Encoding and Transmission Standard, METS). Information about the format (or MIME type) will most often be provided in a *format* element. Metadata for any special equipment, computing environment, or software necessary to reuse the object are also important to provide to the user.

Accessibility and discoverability will also depend on the existence of high-quality metadata. The more you have, and the more organized it is, the easier it will be to search for an object. Users query databases for information and objects based on the metadata that exists for an object. Searching by *author*, *title*, *format*, or a phrase in the *description* requires that information of those kinds exist (a value for each of those fields in a metadata record). Attaching the object to its record (if it is digital) makes it even more accessible, and this can be done with an *identifier*.

The following Table on different metadata standards and their respective functions is a reproduction of one originally created by Anne J. Gilliland for The J. Paul Getty Trust’s *Introduction to Metadata* (Online Edition, Version 3.0 http://www.getty.edu/research/publications/electronic\_publications/intrometadata/setting.html). We recommend reading through the entire document.

|  |  |  |
| --- | --- | --- |
| **Type** | **Definition** | **Examples** |
| Administrative | Metadata used in managing and administering collections and information resources | Acquisition information;  Rights and reproduction tracking;  Documentation of legal access requirements;  Location information;  Selection criteria for digitization |
| Descriptive | Metadata used to identify and describe collections and related information resources | Cataloging records;  Finding aids;  Differentiations between versions;  Specialized indexes;  Curatorial information;  Hyperlinked relationships between resources;  Annotations by creators and users |
| Preservation | Metadata related to the preservation management of collections and information resources | Documentation of physical condition of resources;  Documentation of actions taken to preserve physical and digital versions of resources, e.g., data refreshing and migration;  Documentation of any changes occurring during digitization, or preservation |
| Technical | Metadata related to how a system functions or metadata behaves | Hardware and software documentation;  Technical digitization information, e.g.,  formats, compression ratios, scaling routines;  Tracking of system response times;  Authentication and security data, e.g., encryption keys, passwords |
| Use | Metadata related to the level and type of use of collections and information resources | Circulation records;  Physical and digital exhibition records;  Use and user tracking;  Content reuse and multi-versioning information;  Search logs;  Rights metadata |

**Resources:**

1. National Information Standards Organization (NISO). 2004. Understanding Metadata. <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>
2. Miller, Steven J. 2011. Metadata Resources: Selected Reference Documents, Web Sites, and Readings: <https://pantherfile.uwm.edu/mll/www/resource.html>
3. Wikipedia page on “Metadata”: <http://en.wikipedia.org/wiki/Metadata>
4. University of Illinois at Urbana-Champaign. Best Practices for Structural Metadata

<http://www.library.illinois.edu/dcc/bestpractices/chapter_11_structuralmetadata.html>

**Sample metadata standards**

Adhering to metadata standards is crucial to successful data management and for future publishing and funding. Metadata standards guide the collection and structure of metadata so that data is collected, described, structured, and referred to consistently.

While it’s generally agreed that good metadata is the key to discovering and sharing research data, given the great variety of metadata specifications, deciding which metadata to capture and which standard to use can be difficult for researchers and data curators alike. Many academic communities have agreed upon metadata standards that best meet the needs for reuse of their discipline specific data. The Digital Curation Centre (<http://www.dcc.ac.uk/>) provides a list of these disciplinary metadata standards. A sampling of these standards is provided below as an example.

|  |  |  |
| --- | --- | --- |
| **Discipline** | **Metadata Standard** | **Description** |
| Biology | Darwin Core  <http://www.dcc.ac.uk/resources/metadata-standards/darwin-core> | A body of standards, including a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. |
| Ecology | EML - Ecological Metadata Language  <http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language> | Ecological Metadata Language (EML) is a metadata specification particularly developed for the ecology discipline. |
| Earth Science | AgMES - Agricultural Metadata Element Set  <http://www.dcc.ac.uk/resources/metadata-standards/agmes-agricultural-metadata-element-set> | A semantic standard for description, resource discovery, interoperability and data exchange for different types of agricultural information resources. |

|  |  |  |
| --- | --- | --- |
| Climatology | CF (Climate and Forecast) Metadata Conventions  <http://www.dcc.ac.uk/resources/metadata-standards/cf-climate-and-forecast-metadata-conventions> | A standard for climate and forecast “use metadata” that aims both to distinguish quantities (such as physical description, units, or prior processing) and to locate the data in space–time. |
| Physical Science | CIF - Crystallographic Information Framework  <http://www.dcc.ac.uk/resources/metadata-standards/cif-crystallographic-information-framework> | An extensible standard file format and set of protocols for the exchange of crystallographic and related structured data. |
| Social Sciences & Humanities | DDI - Data Documentation Initiative  <http://www.dcc.ac.uk/resources/metadata-standards/ddi-data-documentation-initiative> | An international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification supports the entire research data lifecycle. |

|  |  |  |
| --- | --- | --- |
| General Research Data | DataCite Metadata Schema  <http://www.dcc.ac.uk/resources/metadata-standards/datacite-metadata-schema> | A domain-agnostic list of core metadata properties chosen for the accurate and consistent identification of data for citation and retrieval purposes. |

|  |  |  |
| --- | --- | --- |
|  | Dublin Core  <http://www.dcc.ac.uk/resources/metadata-standards/dublin-core> | A basic, domain-agnostic standard which can be easily understood and implemented, and as such is one of the best known and most widely used metadata standards. |

**Resources:**

1. Digital Curation Centre’s Disciplinary Metadata resource. <http://www.dcc.ac.uk/resources/metadata-standards>.
2. Hogrefe, K., Stocks, K. 2011. "The Importance of Metadata Standards." *In The MMI Guides: Navigating the World of Marine Metadata*. <http://marinemetadata.org/guides/mdatastandards/stdimportance>. Accessed March 22, 2013.

**Other suggested readings**

1. Introduction to Metadata: Setting the Stage (Getty Research Institute) <http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html>
2. Documentation and Metadata (MIT Libraries): <http://libraries.mit.edu/guides/subjects/data-management/metadata.html>

**Part II**

**Interoperability:**

“… ability of a system or a product to work with other systems or products without special effort on the part of the customer. Interoperability is made possible by the implementation of standards.”

*From the IEEE Standards Glossary*

**Controlled vocabularies and technical standards**

As indicated throughout this chapter, metadata is *structured* information about a resource.   Metadata standards, such as Dublin Core, help organize information by providing general guidance and syntax rules. However, because there has been a proliferation of different metadata standards to meet the research needs for different communities, standards also make use of *controlled vocabularies* and *technical standards* in order to facilitate interoperability. In order to ensure that your information will be of use to other researchers, it is important to be aware of how both concepts help you describe and document your data.

Controlled vocabularies are simply lists of predefined terms that ensure consistency of use, and help to disambiguate similar concepts. It is usually a good idea to use the controlled vocabulary that best matches the type of research you are describing. For example, subject terms used in research about biometric sensing may be taken from a controlled vocabulary list such as the Medical Subject Headings (MeSH) (<http://www.nlm.nih.gov/mesh/>). Controlled vocabularies are important because they solve the problems of natural language ambiguity such as homographs and synonyms. In short, controlled vocabularies ensure consistency and clarity.

**A rose by any other name:**

The Library of Congress Subject Heading for roses is: *Roses*. Straight forward enough, but consider the fact that the LCSH for the common fruit fly is: *Drosophila.* Never assume that a controlled vocabulary will enter its terms according to your preferred usage. Always check first!

For example, the Library of Congress Subject Headings (LCSH) (<http://authorities.loc.gov/>) take the guess work out of choosing between: a preferred spelling (catalog versus catalogue), a scientific or popular term (Parrots versus Psittacidae), or determining which synonym to use (automatons versus robots). Some other examples of controlled vocabularies include the ERIC Thesaurus for education terms (<http://eric.ed.gov/),>the IET INSPEC Thesaurus of the Scientific and Technical terms (<http://www.theiet.org/resources/inspec/products/aids/index.cfm>), and the Centre for Agricultural Bioscience international’s CAB Thesaurus (<http://www.cabi.org/cabthesaurus/mtwdk.exe?yi=home>).

A principle of good metadata is that it uses technical standards to help describe the content. Technical standards ensure that the units such as date and time, format, etc. are entered consistently amongst different researchers. Date and time are particularly troublesome to enter consistently because of different types of notation. Consequently, you may choose to use the World Wide Web Consortium Date and Time Format (W3C-DTF) which provides strict encoding rules about how date information is entered. It is a profile based on another international standard, ISO 8601. This is important because different metadata standards may need different levels of granularity in the date and time and because different communities have different ways of expressing dates. The formats and required punctuation are found below.

**Standard organizations:**

A standards organization is any organization whose primary activities are developing and coordinating [technical standards](http://en.wikipedia.org/wiki/Technical_standard) such as weights and measures, web encoding standards, time, etc… . Some examples include ISO, NISO, IEEE, and W3C.

Year:

YYYY (e.g. 1997)

Year and month:

YYYY-MM (e.g. 1997-07)

Complete date:

YYYY-MM-DD (e.g. 1997-07-16)

Complete date plus hours and minutes:

YYYY-MM-DDThh:mmTZD (e.g. 1997-07-16T19:20+01:00)

Complete date plus hours, minutes and seconds:

YYYY-MM-DDThh:mm:ssTZD (e.g. 1997-07-16T19:20:30+01:00)

Complete date plus hours, minutes, seconds and a decimal fraction of a second

YYYY-MM-DDThh:mm:ss.sTZD (e.g. 1997-07-16T19:20:30.45+01:00)

Note that the "T" appears literally in the string, to indicate the beginning of the time element, as specified in ISO 8601.

By formatting your date elements according to this standard, you not only ensure that a machine can “read” it, but also a colleague from France.

Don’t worry about knowing all the different technical standards and controlled vocabularies. Typically the metadata standard you use will provide a best practice recommendation for which controlled vocabularies and standards you should enter. Consider this entry from the Dublin Core Metadata Initiative (DCMI) for the term: *format*.

|  |  |
| --- | --- |
| **Term Name:  format** | |
| **URI:** | <http://purl.org/dc/terms/format> |
| **Label:** | Format |
| **Definition:** | The file format, physical medium, or dimensions of the resource. |
| **Comment:** | Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME]. |
| **References:** | [MIME] <http://www.iana.org/assignments/media-types/> |

**MIME media types:**

An internet standard that originally extended the formats e-mail could support, and is now used to describe content types in general.

The recommended controlled vocabulary is the Multipurpose Internet Mail Extensions (MIME) media types, and the “Reference” element points to the URI for standard organization (IANA) which maintains the controlled vocabulary for the appropriate formats. If you use the Dublin Core term format, you can choose from the following options:

application, audio, example, image, message, model, multipart, text, or video.

Needless to say, there are standards and controlled vocabularies for every conceivable element you may wish to describe. For instance, ISO 639 provides a set of language codes for representing the language of a resource. Again, your metadata standard will generally recommend a best practice with the idea that as long as you structure your data according to the defined standards, it will be consistent, and able to be discovered and reused by other researchers. In cases where it is unclear, or not defined, it may help to talk to a metadata specialist, who can advise and help with your documentation.

**Metadata elements**

At this point the number of metadata standards, controlled vocabularies and technical standards available to you may seem daunting. It is important to remember that the metadata standards are frequently designed for a specific purpose, which should dovetail with the types of controlled vocabularies and technical standards that best describe your data. In other words, someone in the cultural heritage community may want to use Dublin Core and LCSH whereas a biologist may find Darwin Core and MeSH more appropriate. Overtime you will become more proficient in recognizing the metadata standard for your research community.

**Elements:**

“The individual pieces of information collected about a resource. They often correspond to fields when entering the information into a database or spreadsheet.”

From: *Bibliographic/Multimedia Database Model Documentation (UW Core Metadata Companion)*

*UW Madison Libraries’ Local Usage Guide and Interpretations*

Nonetheless**,** there are some common *elements* necessary to ensure that you data can be found and used by other researchers. The following is taken from MIT’s best practices for managing your data. (See MIT’s website at <http://libraries.mit.edu/guides/subjects/data-management/metadata.html>). These elements are necessary regardless of your discipline, and can be used as a general crib sheet if you are not using an established metadata standard.  At minimum, store this documentation, including the description of each element in a generic text (.txt) file, together with your data.

|  |  |
| --- | --- |
| ***Title*** | *Name of the dataset or research project that produced it* |
| ***Creator*** | *Names and addresses of the organization or people who created the data* |
| ***Identifier*** | *Number used to identify the data, even if it is just an internal project reference number. This should always be a unique number.* |
| ***Subject*** | *Best practice is to use a controlled vocabulary to establish the appropriate keywords or phrases describing the subject or content of the data* |
| ***Funders*** | *Organizations or agencies who funded the research* |
| ***Rights*** | *Any known intellectual property rights held for the data* |
| ***Access information*** | *Where and how your data can be accessed by other researchers* |
| ***Language*** | *Best practice is to use a technical standard to indicate the language(s) of the intellectual content of the resource, when applicable* |
| ***Dates*** | *Best practice is to use a technical standard to indicate key dates associated with the data, including: project start and end date; release date; time period covered by the data; and other dates associated with the data lifespan, e.g., maintenance cycle, update schedule* |
| ***Location*** | *Where the data relates to a physical location, record information about its spatial coverage* |
| ***Methodology*** | *How the data was generated, including equipment or software used, experimental protocol, other things one might include in a lab notebook* |
| ***Data processing*** | *Along the way, record any information on how the data has been altered or processed* |
| ***Sources*** | *Citations to material for data derived from other sources, including details of where the source data is held and how it was accessed* |
| ***List of file names*** | *List of all data files associated with the project, with their names and file extensions (e.g. 'NWPalaceTR.WRL', 'stone.mov'). Best practice is to establish a file naming convention to ensure ease of discoverability* |
| ***File Formats*** | *Format(s) of the data, e.g. FITS, SPSS, HTML, JPEG, and any software required to read the data* |
| ***File structure*** | *Organization of the data file(s) and the layout of the variables, when applicable* |
| ***Variable list*** | *List of variables in the data files, when applicable* |
| ***Code lists*** | *Explanation of codes or abbreviations used in either the file names or the variables in the data files (e.g. '999 indicates a missing value in the data')* |
| ***Versions*** | *Date/time stamp for each file, and use a separate ID for each version* |
| ***Checksums*** | *To test if your file has changed over time* |

**Creating metadata**

Metadata creation generally comes about by manual entry of data, automatic extraction, or a combination of the manual and automatic methods. The manual method occurs when you enter information about your resource into a template, a table, a spreadsheet or some other sort of data entry interface. Typically manual metadata is descriptive in nature. Automatic metadata creation occurs when information about a resource is extracted, as in information about a photograph’s pixel resolution, time and place taken, etc. Generally this type of metadata is technical in nature. Obviously decisions about who will produce the metadata and what method or combination will be used must be considered as part of your overall project plan. What follows below are some general considerations to help you decide how to manage metadata creation. What follows has been adapted from: *Bibliographic/Multimedia Database Model Documentation (UW Core Metadata Companion) UW Madison Libraries’ Local Usage Guide and Interpretations*

**General metadata creation considerations**

At first metadata may seem complicated. It is not. Its entire purpose is to enable the discovery, use, and reuse of your research. This is particularly important in an increasingly online, and linked digital environment. When in doubt always contact a metadata specialist, as they are there to assist you. Here are some best practices as you prepare to create your own metadata to describe your content.

1. Consistent data entry is important. Review your metadata for typos, extraneous punctuation, and any inconsistencies in fielded entry, such as putting an author into a title field.
2. Avoid extraneous punctuation as it can create retrieval issues.
3. Avoid most abbreviations. It is fine to use common or accepted abbreviations (such as "cm" for "centimeters") as long as you document the expectation, and are consistent about it.
4. In general, capitalize the first word (of a title, for example) and proper names (place, personal and corporate names) and subject terms only. Capitalize content in the description field according to normal rules of writing. Do not enter content in all caps except in the case of acronyms.
5. Use templates and macros when possible. It may be that certain data elements will always be the same. In those cases try to automate the entry as it cuts down on errors.
6. Extract pre-existing metadata from your sources whenever possible. Information about pictures and word documents can be embedded within the resource itself and extracted for quick population of templates.
7. Keep a data dictionary of the elements, technical standards, and controlled vocabularies you use in your project.
8. Always use an established metadata standard. Your discipline probably already has a best practices metadata standard specific to your research needs.

**Readings**

1. Neiswender, C. 2010. Introduction to Metadata. In *The MMI Guides: Navigating the World of Marine Metadata*. <http://marinemetadata.org/guides/mdataintro>.

2. Hogrefe, K., and Stocks, K. 2011. The Importance of Metadata Standards. In *The MMI Guides: Navigating the World of Marine Metadata*. <http://marinemetadata.org/guides/mdatastandards/stdimportance>.

3. Getty Research Institute. Introduction to Metadata: Setting the Stage. <http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html>.

4. National Information Standards Organization (NISO). 2004. Understanding Metadata. <http://www.niso.org/publications/press/UnderstandingMetadata.pdf> .

5. Woodbury, D. 2010. What is Metadata.  [http://vimeo.com/3161893](%20http:/vimeo.com/3161893%20).

6. Miller, Steven J. 2011. Metadata Resources: Selected Reference Documents, Web Sites, and Readings: <https://pantherfile.uwm.edu/mll/www/resource.html>.

7. University of Illinois at Urbana-Champaign. Best Practices for Structural Metadata. <http://www.library.illinois.edu/dcc/bestpractices/chapter_11_structuralmetadata.html>.

8. University of Wisconsin. 2007. Bibliographic/Multimedia Database Model Documentation. UW Core Metadata Companion. UW Madison Libraries’ Local Usage Guide and Interpretations. <http://uwdcc.library.wisc.edu/documents/DC_companionv1.3.pdf>.

9. University of Minnesota Libraries. File Naming Conventions. <http://researchdata.wisc.edu/manage-your-data/file-naming-and-versioning/>.

10. Simmons GSLIS. Managing Files. <http://gslis.simmons.edu/tor/01_01_01mgfiles.php>.

11. UK Data Archive. Version Control and Authenticity.  
<http://www.data-archive.ac.uk/create-manage/format/versions>.

12. MIT Libraries. Documentation and Metadata.   
<http://libraries.mit.edu/guides/subjects/data-management/metadata.html>.

13. Riley, J. and Becker, D. 2009. Seeing Standards: A Visualization of the Metadata Universe.   
<http://www.dlib.indiana.edu/~jenlrile/metadatamap/>.

14. Digital Curation Centre. Disciplinary Metadata resource. <http://www.dcc.ac.uk/resources/metadata-standards>.

15. IEEE Standards Glossary <http://www.ieee.org/education_careers/education/standards/standards_glossary.html>

**Sources for this unit**

This unit on metadata consolidates, and makes liberal use of, the following sources:

Controlled vocabularies and technical standards

* <http://en.wikipedia.org/wiki/Controlled_vocabulary>
* <http://www.ieee.org/education_careers/education/standards/standards_glossary.html>

File naming

* [*http://gslis.simmons.edu/tor/01\_01\_01mgfiles.php*](http://gslis.simmons.edu/tor/01_01_01mgfiles.php)
* <https://www.lib.umn.edu/datamanagement/metadata>

Metadata elements

* <http://libraries.mit.edu/guides/subjects/data-management/metadata.html>

Creating metadata

* <http://uwdcc.library.wisc.edu/documents/DC_companionv1.3.pdf>

In addition to the above, you may also wish to consult the following sources:

Version control and authenticity

* <http://data-archive.ac.uk/create-manage/format/versions>

What is Metadata?

* <http://vimeo.com/3161893>

Introduction to Metadata: Setting the Stage

* <http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.pdf>

Seeing Standards: A Visualization of the Metadata Universe

* <http://www.dlib.indiana.edu/~jenlrile/metadatamap/>