Digital Library Curriculum Development
Module 4-b: Metadata

Draft: 6 May 2008

1. **Module name:** Metadata

2. **Scope:** This module addresses uses of metadata and some specific metadata standards that may be appropriate in the context of a DL, and the development of metadata records for digital objects.

3. **Learning objectives:**
    a. Students will be able to:
        i. Explain the basic principles of the design of metadata schema, and the assignment of metadata values for specific materials.
        ii. Design a metadata schema and assign the values appropriate to materials in a particular digital library.

4. **5S characteristics of the module:**
    a. Structures: Metadata is a means for increasing the structure of data objects.
    b. Spaces: Many metadata schemas require the existence of data spaces such as namespaces and repositories, and enable the duplication of data between spaces, as in harvesting.
    c. Scenarios: Different metadata schemas are designed for different use scenarios, such as for administration and preservation.
    d. Societies: Different metadata schemas are designed for different use cases and communities.

5. **Level of effort required:**
    a. In class: 3 hours
    b. Outside of class:
        i. 2-3 hours for readings
        ii. Approximately an hour, for an optional homework assignment (see Exercises, below).

6. **Relationships with other modules:**
    a. 3-b: Digitization: Modules 3-b and 4-b should be taught around the same time in the semester.
    b. 4-d: Subject description, vocabulary control, thesauri, terminologies: Module 4-b is a prerequisite to 4-d.
    c. 8-a: Preservation: Module 4-b is a prerequisite to 8-a.

7. **Prerequisite knowledge required:** Knowledge of the concept of a field, and the difference between structured and unstructured data.

8. **Introductory remedial instruction:**
    a. Introduction
        i. Item-level vs. Collection-level description: Description of an individual data object or a collection of data objects (e.g., a library collection, online repository, etc.)
        ii. Original to the digital library vs. translated to Dublin Core (DC): Creating crosswalks between the metadata native to an object or repository and DC.
        iii. Relationship between a metadata record and an object: One-to-one principle (see below).
        iv. Embedded vs. Associated metadata: Metadata embedded in the data object (e.g., cataloging-in-publication data in books) or separate from the object (e.g., card catalog cards)
        v. Resource discovery: Identifying previously unidentified (to the user) data objects.

    b. Resource Description Framework (RDF)
        i. An expression in RDF is a "triple," consisting of a subject, a predicate and an object. A set of RDF triples is called an RDF graph.
            1. Subject: The object being described (e.g., the sky).
            2. Predicate: An element or field describing the object (e.g., color).
            3. Object: The value that the predicate takes on (e.g., blue).

9. **Body of knowledge:**
    a. Conduct the *Invent object description* exercise prior to addressing Dublin Core. If this exercise is not used, substitute a lecture or other classroom activity to cover the same content.

    b. Metadata:
        i. Simple definition: Data about data
        ii. More detailed definition:
            1. Can refer to either a schema for describing data objects, or the data that describes a specific data object.
                a. The data object can be arbitrarily large or small: a single item (e.g., item-level description) or an entire collection of materials (e.g., collection-level description).

2. A set of structured fields and the values that populate those fields.
3. May be used for a variety of purposes, including discovery, administration, and preservation of data objects.

c. Dublin Core (DC):
    i. DC Principles:
        1. Dumb-Down: Any element may be ignored or simplified in describing an object.
        2. One-to-One: A single metadata record describes a single object.
    ii. Data model:
        1. Machine-processability requires a coherent data model
        2. Provides explicit definitions of resources
        3. Relates DC principles & practices to the developments outside Dublin Core Metadata Initiative (DCMI)
        4. Makes clear the relationship of DC "packages" of information to other metadata "packages"
    iii. Problems with DC:
        1. Minimalism: both a strength and a weakness
        2. Only a few communities have extensions
        3. Most search engines ignore META tags
    iv. DC Extensions: May be developed for specific domains and/or user communities.
        1. Gateway to Educational Materials (GEM): thegateway.org
            a. Example elements include: instructional method, standards, duration, frmat, resources, etc.
        2. Darwin Core: darwincore.calacademy.org
            a. Example elements include: kingdom, phylum, class, order, family, genus, species, scientific name, continent, etc.

d. Namespaces & Repositories
    i. 3 types of interoperability
        1. Federation: Metadata records from multiple repositories are collected into a single repository (e.g., WorldCat). This requires the active participation of the host institutions for the multiple repositories.
        2. Harvesting: Metadata records from one repository are *automatically* gathered and deposited in another repository (e.g., Gateway to Educational Materials). As long records are exposed, they may be harvested without the active participation of the repositories' host institutions.
        3. Crosswalks: "Translation" of records across differing schemas (e.g., OCLC Metadata Crosswalk Repository,

http://www.oclc.org/research/researchworks/schematrans/default.htm).

e. Administrative Metadata
    i. Designates information about the provenance, management of digital objects.
    ii. Should enable verification of the integrity, ownership, and authorship.
    iii. Examples of elements (from the DCMI Administrative Metadata Working Group):
        1. Rights: Information about intellectual property or other rights held in and over the content metadata.
        2. Handling: Instructions for handling the administrative metadata schema and the metadata record.
        3. Affiliation: The organization with which a named person was associated when involved with the resource.

f. Preservation Metadata
    i. See module 8a: Approaches to archiving and repository development
        1. Note: This module is a prerequisite to module 8-a.
    ii. Data necessary to maintain the viability, renderability, and understandability of digital objects over the long term.
        1. Viability: The digital object's bit stream must be intact and readable.
        2. Renderability: The bit stream must be translatable into a form that can be viewed by human users or processed by computers.
        3. Understandability: Contextual information must be provided so that content can be interpreted and understood by users.
    iii. Preservation metadata can be input to preservation processes, and also record the output of preservation processes.
    iv. Examples of elements (from the PREMIS Data Dictionary, v.1.0):
        1. PreservationLevel: A value indicating the set of preservation functions expected to be applied to the object.
        2. Fixity: Information used to verify whether an object has been altered in an undocumented or unauthorized way.
        3. Environment: The means by which the user renders and interacts with content. Separation of digital content from its environmental context can result in the content becoming unusable.

g. Harvesting
    i. Facilitates:
        1. Reuse: Enables uses of data objects by organizations and services other than the owner.
        2. Services: Functionality can be provided to provide web-based services.

3. Communities: Data objects can be shared widely among a community of interest.

ii. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
    1. Harvester: a client application that issues OAI-PMH requests.
    2. Repository: a network accessible server that can process OAI-PMH requests.
    3. Item: a constituent of a repository from which metadata about a resource can be disseminated.
    4. Unique Identifier: unambigiously identifies an item within a repository.
    5. Record: metadata expressed in a single format.
    6. Set: an optional construct for grouping items for the purpose of selective harvesting.

h. Educational metadata
    i. Sharable Content Object Reference Model (SCORM)
        1. Standardizes:
            a. How learning content can be exchanged between systems.
            b. How to define the intended behavior and logic of complex learning experiences so content can be reused, moved, searched for, and recontexualized.
        2. Three parts:
            a. Overview: about the model, vision and future
            b. Content Aggregation Model: how to put learning content together so it can be moved and reused.
            c. Run Time Environment: How content is launched and the learner's progress is tracked and reported back.

i. Semantic Web
    i. "The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects." (quote is from the W3C's Semantic Web Activity website: http://www.w3.org/2001/sw/)
        1. Both of these – common formats and language for representing relationships between real and digital objects – are metadata problems:
            a. Common formats for integration and combination of data may be metadata formats.
            b. Language for representing relationships between objects is an RDF triple.
    ii. Necessitates the widespread use of ontologies
        1. Mapping domain knowledge across many domains

iii. Cyberinfrastructure: Goal of ubiquitous information infrastructure

10. Resources

a. Required readings:
   i. Weibel, Stuart. (1995). Metadata: The Foundations of Resource Description. D-Lib Magazine, 1(1). http://www.dlib.org/dlib/July95/07weibel.html
   ii. Duval, E., Hodgins, W., Sutton, S., Weibel, S. L. (2002). Metadata Principles and Practicalities. D-Lib Magazine, 8(4). http://www.dlib.org/dlib/april02/weibel/04weibel.html

b. Introduction to metadata
   i. Duval, Erik. (2002). Metadata Principles and Practicalities. D-Lib Magazine, 8(4). http://www.dlib.org/dlib/april02/weibel/04weibel.html
   ii. Liddy, Elizabeth. (2005). Metadata: A Promising Solution. EDUCAUSE Review, 40(3), 10–11. http://www.educause.edu/apps/er/erm05/erm0536.asp
   iii. Weibel, Stuart. (1995). Metadata: The Foundations of Resource Description. D-Lib Magazine, 1(1). http://www.dlib.org/dlib/July95/07weibel.html

c. Metadata for DLs
   i. Duval, E., Hodgins, W., Sutton, S., Weibel, S. L. (2002). Metadata Principles and Practicalities. D-Lib Magazine, 8(4). http://www.dlib.org/dlib/april02/weibel/04weibel.html

d. Dublin Core
   i. Dublin Core Metadata Initiative. Using Dublin Core. http://dublincore.org/documents/usageguide/
   ii. Dublin Core Extensions
      1. GEM 2.0: Element Descriptions: http://www.thegateway.org/about/documentation/metadataElements/

e. Namespaces & Repositories
   i. Grace A. (2003). Developing a Metadata Strategy. Cataloging & Classification Quarterly, 36(3/4), 31-46.

f. Administrative Metadata
   i. Metadata Encoding & Transmission Standard (METS)
      1. http://www.loc.gov/standards/mets/
   ii. Text Encoding Initiative (TEI)
      1. Sperberg-McQueen, C. M., Burnard, L. (2002). Guidelines for Text Encoding and Interchange. University of Oxford.

g. Preservation Metadata
    i. Open Archival Information System (OAIS) reference model
        1. Consultative Committee for Space Data Systems. (2002). Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-R-2, Red Book, Issue 2. http://public.ccsds.org/publications/archive/650x0b1.pdf
    ii. PREMIS (PREservation Metadata: Implementation Strategies)
        1. OCLC/RLG PREMIS Working Group. (2005). Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. Dublin, OH: OCLC Online Computer Library Center, Inc. http://www.oclc.org/research/projects/pmwg/premis-final.pdf
        2. OCLC/RLG PREMIS Working Group. (2004). Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community. Dublin, OH: OCLC Online Computer Library Center, Inc. http://www.oclc.org/research/projects/pmwg/surveyreport.pdf
        3. OCLC/RLG Working Group on Preservation Metadata. (2002). Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects. Dublin, OH: OCLC Online Computer Library Center, Inc. http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

h. Metadata for Harvesting
    i. The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html
    ii. Metadata Preparation for the Gateway to Educational Materials. http://www.thegateway.org/about/documentation/metadata-preparation/metaprep/

i. Metadata for specific DLs
    i. National Science Digital Library Metadata Primer: http://metamanagement.comm.nsdlib.org/outline.html
    ii. National Science Digital Library Metadata Registry: http://metadataregistry.org/

j. Semantic Web
    i. Greenberg, J., Stuart Sutton, S., & Campbell, D. G. (2003). Metadata: A Fundamental Component of the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, 29(4), 16-18. http://www.asis.org/Bulletin/Apr-03/greenbergetal.html

ii. Jacob, E. K. (2003). Ontologies and the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, 29(4), 19-22. http://www.asis.org/Bulletin/Apr-03/jacob.html
iii. Miller, E. (2003). Enabling the Semantic Web for Scientific Research And Collaboration. Paper presented at the NSF Post Digital Library Futures Workshop, Chatham, MA. http://www2.sis.pitt.edu/~dlwkshop/paper_millr.html

11. Concept map

12. Exercises / Learning activities

a. Homework to be completed prior to this lesson:
   i. Inventing object description exercise, below, can be conducted as a homework assignment: assigned object description as homework prior to class, then discussion takes place in class.
   ii. Students should bring a brief write-up or notes on their object description to class.
   iii. This should not be a graded assignment, but just a jumping-off point for in-class discussion.

b. Exercise in class (15-20 minutes recommended): Invent object description
   i. Break up class into small groups.
   ii. Hand out an object to each group: a book, a newspaper, a coffee mug, a plant, a Star Wars action figure, whatever the instructor has sitting around his or her office, the more diverse the better.
   iii. Each group should describe their object as completely as possible, so that a user could locate it in a library / museum / archive.
   iv. Each group should provide their descriptions. Identify elements in common between types of objects & elements unique to specific types of objects.
   v. Discuss limitations of specific elements: e.g., books have authors, action figures have manufacturers, plants don't have either.
   vi. Exercise can be repeated using Dublin Core or, if the course in which this module is taught is project-based, whatever metadata schema is used by the project.

c. Discussion questions for the classroom (use at the end of the class session):
   i. If no search engines use the META tag, what's the point of metadata?
      1. Use for discovery is meaningless if it doesn't include discovery by search engines?
      2. Discovery by better tools than currently exist?
      3. Discovery within small collections with homegrown tools?

13. Evaluation of learning outcomes

a. Homework to be completed subsequent to this lesson: Describe a digital object.
    i. This assignment may be performed by students individually or in small groups, though small groups are recommended.
    ii. This assignment assumes that the course in which this module is taught is project-based, in which the students build a small digital library. Students should describe a digital object that will eventually become part of the course project digital library.
    iii. Students should digitize some piece of content relevant to the course project, or take an already-digitized content, and describe it by assigning values to the fields of this metadata schema. (Taking already-digitized content is recommended.)
    iv. Students should select an existing or invent a new metadata schema that can be used in the course project to describe the digital object. (Selecting an existing scheme is recommended.)
    v. Assignment deliverable: Students or groups should:
        1. Select or invent a set of metadata fields appropriate for describing the digital object.
        2. Select or invent values to populate those fields.
        3. Explain their rationale for selecting or inventing those fields and values.
        4. Post their ideas about a metadata schema and field values to an online forum for discussion, and/or discuss in class.

14. Glossary

a. Aboutness: The subject or topic of a document.
b. Cyberinfrastructure: New tools and environments for research that make use of high-performance computing and networking to support collection and management of large data sets.
c. Darwin Core: A metadata standard designed to facilitate the exchange of information about the geographic occurrence of species and the existence of specimens in collections.
d. Encoded Archival Description (EAD): A metadata standard for encoding archival finding aids using Extensible Markup Language (XML).
e. Dublin Core (DC): A shorthand term for the Dublin Core Metadata Element Set, the primary work of the Dublin Core Metadata Initiative (DCMI); a set of 15 metadata elements intended to be generic and universally useful for object description.
f. Extensibility: A characteristic of DC; the ability for elements to be added to a schema as needed for a particular context or use.
g. Gateway to Educational Materials (GEM): A set of metadata elements specific to the domain of education; an extension of DC.
h. Interoperability: The ability of two or more systems or components to exchange information and to use the information that has been exchanged.

i. Metadata: Data about data; structured description of information objects; used to aid the understanding, administration, and use of information objects.
j. Ontology: A data model that represents a set of concepts, and the relationships between those concepts, within a specific domain (as opposed to the definition of the term “ontology” in Philosophy). Often used interchangeably with the term “thesaurus.”
k. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): A set of protocols for interoperability between data repositories.
l. Resource discovery: Identifying previously unidentified (to the user) data objects.
m. Sharable Content Object Reference Model (SCORM): A set of standards for interoperability and reusability of web-based learning materials.
n. Text Encoding Initiative (TEI): A standard that enables libraries, museums, publishers, and individual scholars to represent texts for online research, teaching, and preservation.

15. Useful links

16. Contributors

a. Initial author: Jeffrey P. Pomerantz
b. Evaluators: Jane Greenberg