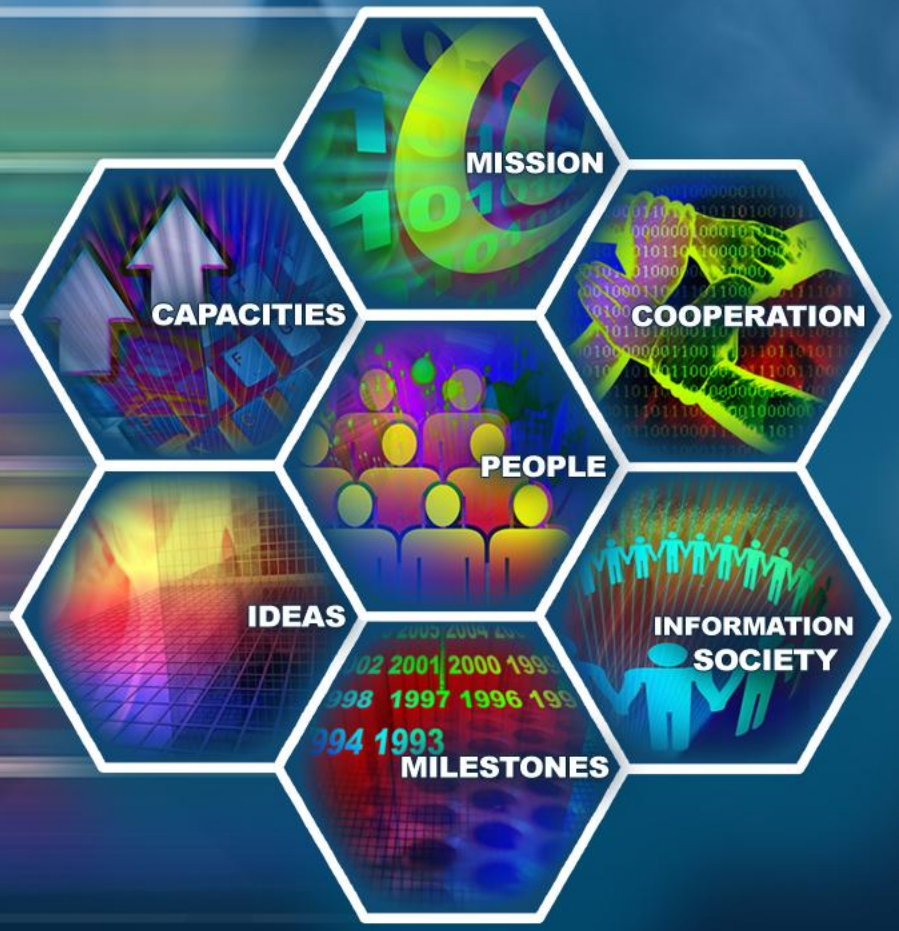
POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER



POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

ACCESS IT Plus – Training workshop in Veria (March ’12)
Session 2: Describing digital objects

Speaker: Marcin Werla (mwerla@man.poznan.pl)

Education and Culture DG
Culture Programme

What is metadata?

- Metadata is

  DATA
  about
  DATA

  - The top data (metadata) is description
  - The bottom data is the object being described
- Metadata is
  - data that operates at a higher level of abstraction
- Example:
  - Data is digital object
  - Metadata is the description of this object
  - Tricky part: what will be the name of the description of the metadata e.g. information about metadata schema?
    - metametadata? ☺
- So: one man's data can be another man's metadata ☺

What is metadata?

- Metadata is structured information, that enables users (and also computer programs) to easily interpret it
- Leonardo da Vinci, Mona Lisa, woman, portrait
  - Such a description makes it hard to interpret the content of the picture, we don't know who the author was (most likely Leonardo but maybe Mona Lisa), and we don't know who is on the painting.
- After adding some categories, things would be much clearer:
  - Creator: Leonardo da Vinci
  - Title: Mona Lisa
  - Subject: woman
  - Type: portrait

What defines a metadata standard?

- Set of elements
  - They can be interrelated
    - e.g. fields and subfields in MARC
    - e.g. general elements and qualifiers (date, creation date, publishing date)
- Rules e.g.
  - Some elements are obligatory
  - Some elements can be repeated
  - Some elements may have only predefined set of values (controlled vocabulary) e.g. type of object, language
  - Some elements may have predefined way of encoding (date, language)
- Guidelines
  - How to choose subject keywords
  - How to determine title (several titles of one object, no title at all etc.)

What is controlled vocabulary?

- In general: a closed list of words which can be modified only under special procedures
- What are the benefits of using it?
  - More consistent metadata – exactly the same value for the same concept
  - Faster creation of metadata – just select value from the existing list

What is controlled vocabulary?

- Example types of vocabularies:
  - Plain list of words
  - Thesaurus – list of words with relations between them e.g.:
    - Broader/narrower term
    - Synonym
    - Related term
    - Preferred term

    What are the benefits for end user in such case?
  - Subject headings list – similar to thesaurus but allows you to combine values in specific order to make description more precise e.g.:
    - Cult of snakes
    - Lithuania

    vs
    - Lithuania – Cult of snakes
  - Classification
    - like thesaurus but with strict hierarchy e.g. each term can have just one broader term

What is interoperabilty in the context of metadata?

- “the ability of your collection to work alongside other collections, either through shared resource discovery services or by contributing your metadata to other collections”

  http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/metadata-standards-and-interoperability/

- The more precise the metadata standard is, the higher is the interoperability
  - Set of elements
  - Encoding of values
  - Controlled vocabularies

Where does metadata live?

- Usually metadata is stored in a database of a digital library system
  - It must be precisely connected with particular digital object
- In some cases the metadata can be embedded in the digital object itself e.g.
  - PDF
  - JPG
  - TIFF

  Such approach is good for long term preservation and files exchange – its harder to loose the connection between data and metadata

How to classify metadata?

- Descriptive metadata
  - Used for discovery and interpretation of an object
- Structural
  - Describes the logical or physical relationships between the parts of a compound object
- Administrative
  - Used for managing the digital object and providing more information about its creation, lifecycle and any constraints governing its use
  - Sometimes divided into
    - Preservation metadata
    - Source metadata
    - Digital provenance metadata
    - Rights metadata

Most important metadata standards?

- DCMES – Dublin Core Metadata Element Set
  - The original objective of the Dublin Core was to define a set of elements that could be used by authors to describe their own Web resources. Faced with a proliferation of electronic resources and the inability of the library profession to catalogue all those resources. The goal was to define a few elements and some simple rules that could be applied by noncatalogers
  - 15 elements: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights
  - DCMES specification describes what is the interpretation of these fields and what information should be kept inside them.
    - Creator holds information about the entity primarily responsible for making the resource
    - Title is a name given to the resource
    - Date is a point or period of time associated with an event in the lifecycle of the resource
    - …

Most important metadata standards?

- Dublin Core Metadata Terms
  - Defines set of qualifiers for the DCMES elements e.g.
    - Types of dates
    - Types of relations
    - Types of coverage
  - Dumb-down principle: an application that does not understand a specific element qualifier should be able to ignore the qualifier and treat the metadata value as if it were an unqualified (broader) element

Most important metadata standards?

- CDWA - Categories for the Description of Works of Arts
  - created to describe cultural objects like museum exhibits
  - 512 categories (which may have sub categories)
  - identifies a set of 35 core tags which should be used as a minimum
    - Object work type, Title, Display creator, Indexing creator, Display measurement, Indexing measurement, Display material/technique, Indexing material/technique, Display State/Edition, Style, Culture, Display Creation Date, Indexing Dates, Location/Repository, Subject, Classifications, Description, Inscriptions, Related works, Rights to work, Record, Resource

Most important metadata standards?

- LIDO - Lightweight Information Describing Objects
  - Created on top of existing museum schemas
  - LIDO is a data exchange schema and should not be used as a basis for collection management.
  - It was created mainly to deliver rich metadata for organization's online collections database and portals.
  - LIDO record is divided into 7 areas :
    - Object Identification — basic information about the object
    - Object classification — information about the type of object
    - Relation — relations of the object to its subject and to other objects
    - Event — events that the object has taken part in, including: Creation, Acquisition, exhibition, etc.
    - Rights Work — information about the rights associated with the object, metadata and the digital surrogate being harvested into the service environment.
    - Record — basic information about the record
    - Resource — basic information about the digital resource being supplied to the service environment (e.g. to museum portal).

Most important metadata standards?

- ISAD
  - Defines general rules for archival description that may be applied irrespective of the form or medium of the archival material.
  - It consists of twenty six elements that may be combined to create the complete description of an archival entity.
- EAD
  - Standardized XML schema which was developed to mark up a data from archival finding aids.
    - Finding aids are something a bit different than usual metadata record (from the librarian point of view) they are much longer, more narrative but also highly structured.
    - Finding aid starts with a description of the collection as a whole, explaining what kind of materials it holds and why they are relevant. It is really hard to imagine a whole biography of an author as a part of metadata record but it fits perfectly into a finding aid.
    - In the end of such an elaborate document a user will find information about how the given content is physically stored, i.e. the number of boxes.

Most important metadata standards?

- MODS
  - Directly derived from MARC 21
  - Designed to simplify MARC descriptions
    - Subset of MARC 21 fields and it uses natural-language based labels for field rather than field numbers as they are used in MARC
    - MODS is encoded in XML, so all metadata records compliant with MODS are also compliant with XML
  - Although compatible, MODS offers some enhancements over traditional MARC 21, including:
    - an optional ID element to facilitate linking at the element level
    - the ability to specify language, script and transliteration scheme at any level

Most important metadata standards?

- METS is not a descriptive metadata standard;
  - designed rather to provide a standard data structure for complex digital library objects
  - XML-based schema, it allows to create one document holding information about:
    - object structure, e.g. list of pages in a book in an appropriate order
    - associated descriptive, (e.g. a DCMES or MODS metadata record) and administration, (e.g. information about digitisation process) metadata
    - links to files representing the content of the object
  - METS document is built out of 7 sections:
    - METS Header — basic information about METS document, e.g. creator,
    - Descriptive Metadata – descriptive metadata (also in a form of links to external metadata records),
    - Administrative Metadata – technical information about how files were created, stored and other information (also Rights information),
    - File section – list of all files which are a part of this object,
    - Structural Map – models a structure of the objects, links files with appropriate metadata records,
    - Structural Links – “allows METS creators to record the nodes in the hierarchy outlined in the Structural Map”,
    - Behaviour – associates executable behaviour with the content stored in METS file.

How does metadata creation look like?

- Identify requirements
  - What is the purpose of your collection? Who will use it? How big will it be? How much time do you have to prepare metadata?
- Investigate best practices in the field
  - Investigate similar digital collections created by other institutions. Ask questions: why was it done in this way? Why not this or that standard?

How does metadata creation look like?

- Choose/Create a metadata schema and additional resources (e.g. controlled vocabularies)
  - Take a look at requirements, consider solutions used by others, investigate potential metadata standards and additional resources like controlled vocabularies or subject headings.
  - You may also consider creating an application profile or extension for an existing metadata standard, and this would be covered in more details shortly.
- Learn how to use chosen metadata tools
  - You will probably learn while using them but if there is such a possibility, it is very important to confront your knowledge with other professionals who have practical experience with a given schema.
  - If someone else is going to work with the chosen schema, you should tell them how it should be used, and together you can prepare internal metadata creation guidelines.

How does metadata creation look like?

- Create or import metadata
  - Metadata creation
    - Decide how much time you can spend on describing one resource
    - Investigate the object which will be the subject of your work, identify sources of information, purse necessary knowledge
    - Prepare a description for the given object
    - Store the description in the metadata registry/files/digital library.
  - Metadata import
    - If the given object has a description prepared by someone else, try to use their work. You can import metadata from the existing library catalogue or some other source.
    - Remember to investigate whether the imported description is fully compliant with your assumption and metadata schema of your choice.

How does metadata creation look like?

- Evaluate metadata quality
  - If the imported description was associated with the physical object, review it in order to avoid misconceptions.
  - Define criteria of correctness, validate metadata with external validator if available
- Adjust metadata over time
  - When it becomes apparent who are the users of your collection, it might be appropriate to adjust your metadata in order to increase the usage of resources.
    - E.g. look at query logs
  - It might also be necessary to adjust metadata in order to participate in some broader initiatives like Europeana.

Who creates metadata?

- What skills might be useful while creating metadata?
  - Apart from knowledge about cataloguing, it is good to have some domain knowledge.
  - In scientific libraries it is a common practice to involve an author or a domain expert to get some knowledge about the given resource/publication.
  - Such an external help is only the source of knowledge in the field; still the cataloguer is responsible for creating a metadata description.
  - If expert is not available, the cataloguer may try to pursue information about resource on their own but this requires a lot of time.
  - So, if you do not have time you should focus on things which can be done, and a general description is still more useful than none.

What metadata schema will be good for me?

- Who will be using the collection?
  - Will it be an ordinary Internet user or an expert historian?
  - This determines several things: language used in descriptions, a question if it is useful to use controlled vocabularies and how detailed descriptive metadata should be.
- Who is the collection cataloguer?
  - Does the cataloguer have domain knowledge or not?
  - Is he able to fill in the complex metadata schema?
- How much time/money do you have?
  - The richer the object description is, the bigger chance there is that users will use the collection
  - The richer the object description is, the more time (and money) is consumed for its preparation

Which metadata schema will be good for me?

- How will your collection be accessed?
  - The most important issue here are the capabilities of the digital library system which will be used to assure access to your digital resources.
    - Does it support complex metadata schemas?
    - What searching capabilities does it have?
    - If users are not able to use the whole richness of our description, the question is whether it is worth creating such rich metadata.
- How is your collection related to other collections?
  - If you have identified a similar collection, it is very likely that you might be interested in creating a reference to a resource from that collection.
  - It is also very likely that at some point both collections will be searched through the same searching interface
    - In such a case it is worth assuring common understanding of some terms and common encoding rules, e.g. for dates.
  - If your library has an OPAC, it is also a common practice to link between an OPAC record of a physical object and digital reproduction in a digital library.

Which metadata schema will be good for me?

- What is the scope of your collection?
  - By determining the scope you would be able to cut off some non-related metadata standards.
    - Is there any dedicated metadata standard for your type of resources?
    - Are there any controlled vocabularies dedicated to your objects?
  - If a collection with a similar scope exists, it is worth taking a look at solutions which were employed by its developers.
    - But since every digitisation project is different, you should still evaluate their findings on your own to avoid their mistakes.

Which metadata schema will be good for me?

- Will your metadata be harvested?
  - Harvesting is an act of collecting web resources by a dedicated computer program.
  - In the domain of digital libraries special harvesting protocols were developed — Open Archive Protocol for Metadata Harvesting (OAI-PMH) is the most widely used.
  - In case of this protocol collection's metadata must be available at least in DCMES.
    - Your collection may be described using any metadata standard but in order to be an OAI-PMH compliant repository it has to expose metadata also in Dublin Core.
  - If DCMES seems too simple for your collection and you will choose a more suitable standard, you will have to remember to provide an appropriate mappings and conversion mechanism to enable OAI-PMH harvesting of DCMES.

What is an application profile?

- Various metadata standards define mechanisms which allow for schema extensions - so called application profiles (AP).
  - It is defined as “schemas which consist of data elements drawn from one or more namespaces combined together, and optimised for a particular local application”
- Creating an application profile is usually associated with mixing various metadata standards in order to provide better resource discovery within a given domain.
- It is always better to qualify an existing element rather than create a new additional element.
  - It is possible to collapse richer representation into the original schema
- When adding new elements, it is preferable to draw on other metadata standards if possible.
- You should document introduced changes publicly.
  - This would allow interested parties to understand the applied approach and use digital resources from the collection more efficiently.

How to evaluate metadata quality?

- Evaluation needs to consider different layers of metadata record correctness
  - its compliance to the chosen standard,
  - its accuracy and correctness of the description in terms of domain knowledge,
  - and whether metadata record serves to its purposes.
- Metadata record should contain as much information as necessary to answer four questions:
  - who? what? where? when?

  These four high level entry points should be present in any descriptive metadata

Do I really have to prepare internal metadata guidelines?

- It will allow you to clearly describe why you decided to use a given standard,
  - If you have created your own application profile, you should explain and document your approach.
    - Which fields are mandatory (if any),
    - Which elements were added to the schema and for what purpose.
  - Which vocabularies were chosen for each element of description.
  - What encoding rules were applied for dates, time periods, proper names, places, etc.
  - Criteria used to evaluate the quality of the created metadata
- Thanks to this you will be able to share information about your experience with other professionals
  - Other people would be able to contribute to your work.
- Such a document, when publicly available might also be helpful for your users while they are searching and browsing the collection.
- After some time you will most likely forget why you decided to use one solution over the other.
  - Internal metadata guidelines would be the great foundation for advancing development of your present and future collections.

Principles aimed to facilitate creation of 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.

Principles aimed to facilitate creation of good metadata

- 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.

Metadata for digitized objects

- One-to-one rule: One metadata record can describe only one resource.
  - You should not combine information about physical features of the original object and its scanned version.
  - As the intellectual content is the same, other features of these resources are different, e.g. the carrier: oil paiting canvas and a digital image stored in the JPG format
    - Format
      - Physical object: weight and dimension of the painting,
      - Digital image: size (weight) and resolution of the JPG file,
    - Type
      - Physical object: an oil painting,
      - Digital version: a JPG image,
    - Date
      - Physical object: when the painting was created,
      - Digital version: when the digital reproduction was created,
    - Creator
      - Physical object: Famous painter,
      - Digital version: Person who created the digital image.
  - In an ideal case you should maintain two DCMES records:
    - one for the original object
    - one for the digital copy

Introducing Dublin Core Terms

- Title
  - Alternative
- Creator
- Subject
- Description
  - Abstract
  - TableOfContents
- Publisher
- Contributor

Introducing Dublin Core Terms

- Date
  - Available
  - Created
  - DateAccepted
  - DateCopytrighted
  - DateSubmitted
  - Issued
  - Modified
  - Valid
- Type
- Format
  - Extent
  - Medium

Introducing Dublin Core Terms

- Identifier
  - BibliographicCitation
- Source
- Language

Introducing Dublin Core Terms

- Relation
  - ConformsTo
  - HasFormat
  - HasPart
  - HasVersion
  - IsFormatOf
  - IsPartOf
  - IsReferencedBy
  - IsReplacedBy
  - IsRequiredBy
  - IsVersionOf
  - References
  - Replaces
  - Requires

Introducing Dublin Core Terms

- Coverage
  - Spatial
  - Temporal
- RightsHolder
- Rights
  - AccessRights
  - License

Introducing Dublin Core Terms

- AccuralMethod
- AccuralPeriodicity
- AccuralPolicy
- Provenance
- Audience
  - EducationLevel
  - Mediator
- InstructionalMethod

Using Dublin Core terms

- Please refer to the e-learning course lesson **What should I know about Dublin Core Terms?**
  - You will find there a set of example guidelines for several DC elements

POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

End of session 2:
Describing digital objects

Poznań Supercomputing and Networking Center
affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences,
ul. Noskowskiego 12/14, 61-704 Poznań, POLAND,
Office: phone center: (+48 61) 858-20-00,
fax: (+48 61) 852-59-54,
e-mail: office@man.poznan.pl, http://www.man.poznan.pl