Access IT Training

Part 4.
How to describe digital objects?

How to describe digital objects?

- Well… with metadata?
  - Metadata (literally data about data)
    - is a structured data about resources (digital and non-digital)
    - can be used to support a wide range of operations on those resources, like:
      - discovery
      - resource management (including rights management)
      - long-term preservation
  - Different metadata may be required to support different functions

What have to be described?

- During a digitisation project it may be necessary to provide metadata describing several classes of resources including
  - The physical objects digitised
    - cataloguing before digitisation
  - The digital objects created during the digitisation process
    - "digital masters"
  - The digital objects derived from these "digital masters"
    - presentation form of digital objects
  - Collections of any of the above

Metadata classification

- Classification of the metadata standards
  - By source domain, e.g.:
    - Museums – SPECTRUM, CDWA, LIDO
    - Archives – ISAD, ISAAR, EAD
    - Libraries – MARC
  - By the functions that are intended to be supported by the metadata
    - Note: in practice metadata schemas often support multiple functions and overlap categories of such classification

Metadata classification

- Descriptive metadata – used for discovery and interpretation of the digital object
  - The absolute minimum and a must is the Dublin Core Metadata Element Set (DCMES) in its simple/unqualified form
  - 15 basic elements
  - Cross-domain
  - http://dublincore.org/documents/dces/
  - DCMES is:
    - IETF RFC 5013
    - ANSI/NISO Standard Z39.85-2007
    - ISO Standard 15836:2009.

Metadata classification

- Descriptive metadata
  - Review existing metadata models and standard before creating your own
    - Ask similar institutions in your country what they use – it may be also good for you
    - A metadata model once established in your institution may/will be used in many further digitisation projects
    - Creating totally new metadata schema should be avoided
  - If you do not have a very good reason to do so, the Dublin Core should be a basic frame for a metadata schema
    - More about this later…
  - If a proprietary metadata model is to be used, the mapping to Dublin Core should be developed

Metadata classification

- Administrative metadata – used for managing the digital object and providing more information about its creation and any constraints governing its use
  - Technical metadata
    - Describing the technical characteristics of a digital resource
    - Includes information that can be captured effectively only during the digitisation process
    - Some of this information may be generated automatically
  - Source metadata
    - Describing the object from which the digital resource was produced
  - Digital provenance metadata
    - Describing the history of the operations performed on a digital object since its creation
  - Rights management metadata
    - Describing copyright, use restrictions and license agreements that constrain the use of the resource

Metadata classification

- Administrative metadata
  - There is no single standard for such type of metadata
  - A lot of such metadata is stored internally by the software used for management of digital objects

Metadata classification

- Preservation metadata
  - Metadata intended to support preservation
  - PREMIS data dictionary - initially developed with the goal of creating an implementable set of "core" preservation metadata elements, with broad applicability within the digital preservation community
    - http://www.loc.gov/standards/premis/

Metadata classification

- Structural metadata
  - Describes the logical or physical relationships between the parts of a compound object
  - METS provides an encoding format for descriptive, administrative and structural metadata
    - It is designed to support both the management of digital objects and the delivery and exchange of digital objects across systems
    - http://www.loc.gov/standards/mets/

Metadata classification

- Collection-level description
  - Collections are seen as components around which many different types of digital services might be constructed
  - They should be described so that the user can discover important characteristics of the collection including scope, format, ownership and restrictions on access
  - The description also allows collections to be integrated into digital services operating across these collections
  - Dublin Core Collection Description Application Profile can be used for this
    - http://dublincore.org/groups/collections/collection-application-profile/

Terminology standards

- The effective exchange of information stored in metadata records requires:
  - A shared understanding of the metadata schema in use
  - A shared understanding of the terms used as values in the metadata elements
    - This is achieved by
      - adoption of common terminologies
      - establishing of relationships between terms in different terminologies

Terminology standards

- If possible, projects should use multilingual terminological sources
- Local terminologies may be also considered
  - Information about such terminologies should be publicly available
- The use of terminology must be indicated unambiguously in the metadata records

Descriptive metadata

- Dublin Core Metadata Element Set is now a minimum standard in the metadata exchange/interoperability process
- It has two variants
  - DC Simple / DC Unqualified / DC MES
    - 15 elements
    - http://dublincore.org/documents/dces/
  - DC Terms / DC Qualified / DCMI Terms
    - 55 elements
    - http://dublincore.org/documents/dcmi-terms/

Descriptive metadata

- DC Simple
  - contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type

Descriptive metadata

- DC Terms
  - abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, contributor, coverage, created, creator, date, dateAccepted, dateCopyrighted, dateSubmitted, description, educationLevel, extent, format, hasFormat, hasPart, hasVersion, identifier, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, language, license, mediator, medium, modified, provenance, publisher, references, relation, replaces, requires, rights, rightsHolder, source, spatial, subject, tableOfContents, temporal, title, type, valid

Descriptive metadata

- What to choose?
  - Take all 15 DC Simple elements
  - Review DC Qualifiers and take what you find useful in a broad and long-term perspective
  - If you have anything else to add, which is not covered by qualifiers, add your own element
    - As an extra qualifier to one of DC elements

Descriptive metadata

- The official definitions of DC elements are very general
- They can be specified more precisely by documents called application profiles
  - To see what is important when creating such profile see:

    http://dublincore.org/documents/profile-review-criteria/
- Preparation of a document precisely describing how each metadata element should be interpreted and filled is very important

Descriptive metadata

- Such metadata creation manual should contain
  - The list of metadata elements
  - Precise description of each element
  - Examples of complete metadata records for different types of objects
  - Contact details to manual authors/helpdesk allowing the metadata editor to consult in case of doubts

Descriptive metadata

- Metadata creation manual - for each element:
  - Name and general definition
  - Reference to proper DC Term (if an element is not the DC Term itself)
  - Is it mandatory, recommended or optional etc.
  - If possible, the list of digital library functions/services that exploit this element
  - If possible, the places in the digital library interface, where this element is visible
  - Guidelines
  - Examples

Descriptive metadata

- Example: How Europenana defines the importance of DC elements?

Strongly recommended	Recommended	Additional elements
dc:title	dc:coverage	dc:format
dcterms:alternative	dcterms:spatial	dcterms:extent
dc:creator	dcterms:temporal	dcterms:medium
dc:contributor	dc:description	dc:identifier
dc:date	dcterms:isPartOf	dc:rights
dcterms:created	dc:language	dcterms:provenance
dcterms:issued	dc:publisher	dc:relation
	dc:source	dcterms:conformsTo
	dc:subject	dcterms:hasFormat
	dc:type[7]	dcterms:isFormatOf
		dcterms:hasVersion
		dcterms:isVersionOf
		dcterms:hasPart
		dcterms:isReferencedBy
		dcterms:references
		dcterms:isReplacedBy
		dcterms:replaces
		dcterms:isRequiredBy
		dcterms:requires
		dcterms:tableOfContents

Descriptive metadata

- Metadata creation manual - for each element guidelines on:
  - Interpretation in the context of particular resource type
    - E.g. "For printed documents the 'extent' element should contain information about the number of pages and paper format /size"
    - Information like "Do not confuse with" may be useful for problematic elements
  - Allowed values/encoding schemes
    - Closed list of values
      - E.g. "The type element can be only one of TEXT, IMAGE, AUDIO, VIDEO"
    - Reference to external standards
      - E.g. "The language element should be entered according to the ISO 639-2 standard"
    - Reference to external vocabularies
      - E.g. "The subject element should contain at least one element coming from the MeSH vocabulary"
    - General guidelines on the creation of new values
      - "It is recommended to use a standardized writing style for names, so use the writing style used by the publisher in the first place. When that is not applicable use the APA bibliographic writing style as in a reference list when applicable."

DC Simple and selected qualifiers

- Contributor
  - Definition: An entity responsible for making contributions to the resource.
  - Comment: Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.

DC Simple and selected qualifiers

- Contributor
  - Example guidelines
    - Name and surname in reverse order
    - Here you can put the promoter/supervisor of PhD theses
    - Here you can put also patrons, benefactors and sponsors

DC Simple and selected qualifiers

- Coverage
  - Definition: The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
  - Comment: Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.
  - References: [TGN] http://www.getty.edu/research/tools/vocabulary/tgn/index.html

DC Simple and selected qualifiers

- Coverage
  - Qualifiers
    - Temporal
    - Spatial

DC Simple and selected qualifiers

- Creator
  - Definition: An entity primarily responsible for making the resource.
  - Comment: Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.

DC Simple and selected qualifiers

- Creator
  - Example guidelines
    - Use reverse order (surname, initials, prefix)
    - Use separate values for several authors
    - Skip scientific titles
    - Creators with less responsibility can be moved to contributors
    - If not sure, put persons in Creator and institutions in Publisher
    - Use original names of institutions
    - Values like “Mentaly ill patient” are also allowed

DC Simple and selected qualifiers

- Date
  - Definition: A point or period of time associated with an event in the lifecycle of the resource.
  - Comment: Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].
  - References: [W3CDTF] http://www.w3.org/TR/NOTE-datetime

DC Simple and selected qualifiers

- Date
  - Qualifiers
    - Created
    - Accepted
    - Copyrighted
    - Issued
    - Valid
    - Available
    - Modified

DC Simple and selected qualifiers

- Description
  - Definition: An account of the resource.
  - Comment:Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.

DC Simple and selected qualifiers

- Description
  - Qualifiers
    - Abstract
    - Table of Contents
  - Example guidelines
    - Do not confuse with Format

DC Simple and selected qualifiers

- 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/

DC Simple and selected qualifiers

- Format
  - Qualifiers
    - Medium
      - The material or physical carrier of the resource (wood, metal, etc.)
    - Extent
      - The size or duration of the resource (13 cm, 24 minutes etc.)

DC Simple and selected qualifiers

- Identifier
  - Definition: An unambiguous reference to the resource within a given context.
  - Comment: Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.

DC Simple and selected qualifiers

- Identifier
  - Example guidelines
    - You can use systems like URI, URL, DOI, ISBN, ISSN etc. but the first identifier always should be the URL allowing to access the resource

DC Simple and selected qualifiers

- Language
  - Definition: A language of the resource.
  - Comment: Recommended best practice is to use a controlled vocabulary such as RFC 4646 [RFC4646].
  - References: [RFC4646] http://www.ietf.org/rfc/rfc4646.txt

DC Simple and selected qualifiers

- Language
  - Example guidelines
    - This is the language of the resource and not the language of the metadata or language of the title
    - In case of objects without any text or speech this element can be ignored
    - In case of several languages , provide separate value for each language

DC Simple and selected qualifiers

- Publisher
  - Definition: An entity responsible for making the resource available.
  - Comment: Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.

DC Simple and selected qualifiers

- Publisher
  - Example guidelines
    - Use the commercial publisher name and not the organization connected with the creator
    - In case of institutions, remain the hierarchy of units
    - A place of publishing can also go here

DC Simple and selected qualifiers

- Relation
  - Definition: A related resource.
  - Comment: Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.

DC Simple and selected qualifiers

- Relation
  - Qualifiers
    - Conforms to
    - Has Format
    - Has Part
    - Has Version
    - Is Format of
    - Is Part of
    - Is Referenced by
    - Is Replaced by
    - Is Required by
    - Is Version of
    - References
    - Replaces
    - Requires

DC Simple and selected qualifiers

- Rights
  - Definition: Information about rights held in and over the resource.
  - Comment: Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.

DC Simple and selected qualifiers

- Rights
  - Qualifiers
    - Access rights
    - License
    - Rights holder

DC Simple and selected qualifiers

- Rights
  - Example guidelines
    - If there is no such element in the metadata you cannot assume that this is public domain
    - It is recommended to link to a service where the rights are explained in details
    - Metadata should clearly indicate under what terms and conditions / under which license metadata and content can be reused by third parties
  - Examples:
    - Creative Commons Attribution 3.0 License;
    - (c) University of Bath, 2003;
    - http://creativecommons.org/licenses/bysa/2.0/uk;
    - cc-by-sa, Andrew Smith;
    - Licensed for private home exhibition only, any public performance, copying, or other use is strictly prohibited.

DC Simple and selected qualifiers

- Source
  - Definition: A related resource from which the described resource is derived.
  - Comment: The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.

DC Simple and selected qualifiers

- Source
  - Example guidelines
    - Put a location of original object and the telephone number here

DC Simple and selected qualifiers

- Subject
  - Definition: The topic of the resource.
  - Comment: Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. To describe the spatial or temporal topic of the resource, use the Coverage element.

DC Simple and selected qualifiers

- Subject
  - Example guidelines
    - If the subject is person or institution, use the same form as for Creator element
    - If you use formal classification identifiers, add also the descriptive form (l01.451.617 is "Postal Service" in MeSH)
    - Do not confuse with Coverage

DC Simple and selected qualifiers

- Title
  - Definition: A name given to the resource.
  - Comment: Typically, a Title will be a name by which the resource is formally known.

DC Simple and selected qualifiers

- Title
  - Example guidelines
    - Use original title
    - Enter as long title as necessary
    - It should not be confused with the file name
  - Qualifiers
    - Alternative title
      - Example guidelines – it can be:
        - Short form of the title
        - Long form of the title expressed with abbreviation or number
        - Parallel title
        - Correct version of title if there is a mistake in the original title
        - Modern version of an old title
        - Commonly known title
        - Any other form of the title increasing the chance of the object discovery
      - If not used, the above information may be held in the Description element

DC Simple and selected qualifiers

- Type
  - Definition: The nature or genre of the resource.
  - Comment: Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.
  - References: [DCMITYPE] http://dublincore.org/documents/dcmi-type-vocabulary/

DC Simple and selected qualifiers

- Type
  - Example guidelines
    - Do not confuse with format, Type is about the contents of the object and not about physical form.

Descriptive metadata

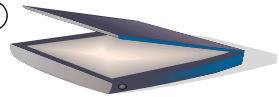
Stone monument, carved, built as a tribute to…

monument 10m x 3m x 3m

What description?

Monument or person? Who is the author?

picture 10cm x 15cm



What size?

image/jpeg 1200px x 1600px

Descriptive metadata

- The most common problem – confusion between original physical object and the digital representation
  - Original object
    - Type: statue
    - Format – size: 10m x 3m x 3m
    - Format – medium: marble stone
  - Digital object
    - Type: photograph
    - Format – size: 1600x1200 px; 1.3 MB
    - Format – medium: image/jpeg
  - The part of the description of the digital representation above can be identical for many different original objects

Descriptive metadata

- Original object
  - Type: statue
  - Format – size: 10m x 3m x 3m
  - Format – medium: marble stone
- Digital object
  - Type: photograph
  - Format – size: 1600x1200 px; 1.3 MB
  - Format – medium: image/jpeg
- Possible solutions:
  - Can we remove the information about digital object? No…
  - Can we remove the information about original object and add link to a source object metadata in our cataloguing system? You can, but is your system on-line? Is it indexed by search engines? Is it as popular as your digital library? Is there any link from the record in this system to the digital object?
  - Can we mix the metadata of both objects? You should not, but this is very common solution…

Descriptive metadata

- Original object
  - Type: statue
  - Format – size: 10m x 3m x 3m
  - Format – medium: marble stone
- Digital object
  - Type: photograph
  - Format – size: 1600x1200 px; 1.3 MB
  - Format – medium: image/jpeg

- Possible solutions:
  - Ideally you could try to maintain both information in parallel with visible distinction in the user interface
    - This maximizes the possibility of object discovery and the possibility of creation of new services
  - Whatever you decide to do, please do it consistently and have end users (and the role of the descriptive metadata for them) in mind

Descriptive metadata

- Description vs Format – example guidelines
  - If data includes information about the craftsmanship or technology related to the object (i.e. pressing, binding, carving, shoe making etc), the preferred is **dc:description.**
  - If data includes information about the physical materials of the object (i.e. ivory, wooden, cast-iron etc), the preferred is **dcterms:medium**
  - If data includes information about both two cases described above, it is preferably **dc:description**

Descriptive metadata

- Where to put spatial information? – example guidelines
  - where an image depicts a place, typically a landscape painting or photograph, use <dc:coverage> or <dc:spatial>
  - where the place depicted is incidental to the story of the object depicted, use <dc:subject>
  - where the relationship is not obvious use <dc:description>
  - where the emphasis of the semantic is on the history of custody and ownership, use <dcterms:provenance>

Descriptive metadata

- DC.Culture
  - An approach developed in the Minerva project, exploits the idea of four high level DC Simple access points:
    - Who?
    - What?
    - Where?
    - When?

DC Culture High-Level element	corresponding DC Simple element(s)
who	creator, publisher, contributor, rights, subject
what	title, subject, description, type, format, identifier, source, language, relation, subject
where	subject, coverage
when	subject, coverage, date

Descriptive metadata

- DC.Culture
  - It is good to prepare this metadata having these four natural access points in mind

DC Culture High-Level element	corresponding DC Simple element(s)
who	creator, publisher, contributor, rights, subject
what	title, subject, description, type, format, identifier, source, language, relation, subject
where	subject, coverage
when	subject, coverage, date

Some bad experience from Poland

- Initially there was no cooperation on the interpretation of DC Simple Schema
  - Now we have dozen of different DC Schemas and more or less the same number of guidelines…

Some bad experience from Poland

- Examples of additions to the schema
  - Description
    - Remarks
  - Publisher
    - Publisher's website
    - Place of publishing
  - Format
    - Physical description
  - Source
    - Number of archival document
    - Name of archival document
  - Relation
    - Full description in the catalogue
    - Series
  - Other
    - Location of original, Digitization, Publication, Digitisation sponsor

Some bad experience from Poland

- The Polish version of the ISO standard with DC Simple had obvious (and serious) translation mistakes…
  - … which were blindly accepted…
    - E.g.: If the Rights element is not present in the metadata you can~~not~~ assume that there are no rights for the described object

Q&A