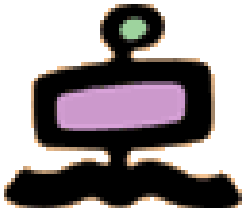
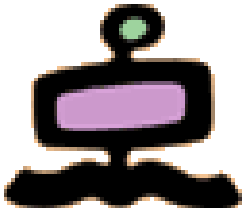
Corso di Biblioteche Digitali



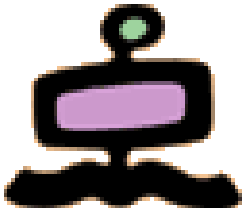
- Vittore Casarosa
  - Mail: casarosa@isti.cnr.it
  - Tel: 050 621 3115 (office) 348 397 2168 (mobile)
  - Skype: vittore1201
- “Ricevimento” at the end of the lessons or by appointment
- Final assessment
  - 70% oral examination
  - 30% project (development of a small digital library))
- Reference material:
  - Ian Witten, David Bainbridge, David Nichols, How to build a Digital Library, Morgan Kaufmann, 2010, ISBN 978-0-12-374857-7 (Second edition)
  - Material provided by the teacher
- **http://cloudone.isti.cnr.it/casarosa/BDG/**

Modules



- Computer Fundamentals and Networking
- A conceptual model for Digital Libraries
- Bibliographic records and metadata
- Information Retrieval and Search Engines
- Knowledge representation
- Digital Libraries and the Web
- Hands-on laboratory: the Greenstone system

Parallel evolution



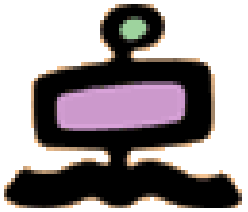
Libraries

- Description (documents)
  - Bibliographic records
  - MARC
- Interoperability
  - Z39-50
- Conceptual model (classes)
  - FRBR – LRM for Works, Expr., Manif.
- Information Retrieval
  - Full text (catalogue and documents)

The Web

- Description (instances)
  - **Metadata**
  - **Dublin Core**
- Interoperability
  - OAI-PMH
- Conceptual model (classes)
  - RDF and RDF Schema for all resources (ontologies)
- Information Retrieval
  - Full text (web pages and resources)

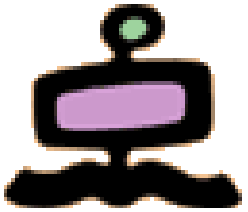
Description of “content”



- In the libraries: bibliographic records
  - Classification and cataloguing
  - MARC standard

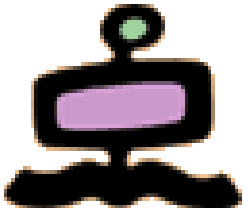
- In the Web: metadata
  - Resources and metadata
  - The Dublin Core metadata schema

Emerging new requirements (early 90's)



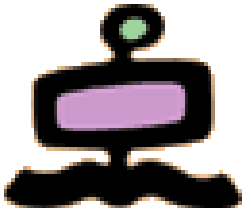
- Increase in the amount of information available on-line (data bases, repositories, the Web, etc)
- Increase in the variety of information available on-line (text, sound, images, video, 3D, etc)
- Scholarly publishing (open access and non-open access)
  - Self-publishing
- Need to describe (in some way) the "content" of the Web
  - Description of information not always done by "specialists"
- Description of the content of the Web done through metadata

Academic publications cycle (before the Web)



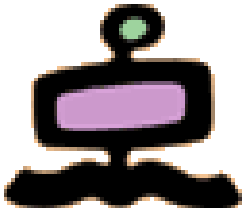
- A paper with “important” scientific results is sent to a scientific journal for publication
- Based on the outcome of a peer review the paper is accepted for publication, usually subject to some revision (and subject to the transfer of the copyright)
- The author sends the revised paper to the journal
- The journal publishes it in the first available issue
  - Usually from one to two years after the initial submission
- With the arrival of the Web, authors immediately saw the possibility of self-publishing on the web the initial version of the paper
- Institutions reacted establishing Institutional Repositories

Institutional Repositories



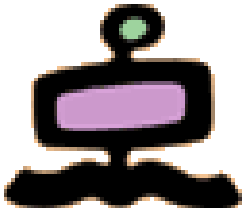
- An Institutional repository is a centrally managed collection of institutionally generated digital objects designed to be maintained “for ever”
- Established and maintained by universities and research institutions (initially) to cope with self-publishing
- An *e-print* is an author self-archived document. The content of an e-print is usually the result of scientific or other scholarly research.
- Repositories contain scholarly publications
  - Reports
  - Working papers
  - Pre- and post-prints of articles and books
  - Doctoral thesis
  - Data supporting research
  - References and professional databases related to research topics
- **The content of an Institutional repository is accessible on the Web**

Advantages of Institutional Repositories



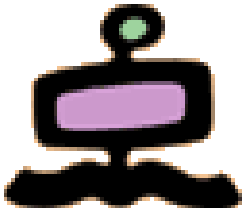
- Opening up outputs of the institution to a worldwide audience;
- Maximizing the visibility and impact of these outputs as a result;
- Showcasing the institution to interested constituencies – prospective staff, prospective students and other stakeholders;
- Collecting and curating digital output;
- Managing and measuring research and teaching activities;
- Providing a workspace for work-in-progress, and for collaborative or large-scale projects;
- Enabling and encouraging interdisciplinary approaches to research;
- Facilitating the development and sharing of digital teaching materials and aids, and
- Supporting student endeavours, providing access to theses and dissertations and a location for the development of e-portfolios.

Emerging new requirements (early 90's)



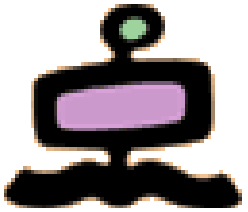
- Increase in the amount of information available on-line (data bases, repositories, the Web, etc)
- Increase in the variety of information available on-line (text, sound, images, video, 3D, etc)
- Scholarly publishing (open access and non-open access)
  - Self-publishing
- Need to describe (in some way) the "content" of the Web
  - Description of information not always done by "specialists"
- Description of the content of the Web done through metadata

Different use of the term “metadata”



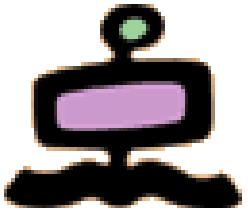
- Database Management Systems (schemas of relational databases, since early ‘70s): machine understandable information
- World Wide Web (since the mid ‘90s): records that describe “resources”
- Library environment: any formal scheme of “document” description, applying to any type of object, digital or non-digital
  - therefore, traditional library cataloging is a form of metadata; MARC 21 and the rule sets used with it, such as AACR2R, are metadata standards
  - other metadata schemes have been developed to describe various types of textual and non-textual objects including published books, electronic documents, archival finding aids, art objects, educational and training materials, scientific datasets etc.

Some definitions of Metadata



- Machine-understandable information about Web resources or other things (Tim Berners-Lee, 1997)
- Data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics; a user might be a program or a person (Lorcan Dempsey, 1998)
- Structured data about resources that can be used to help support a wide range of operations (Michael Day, 2001)
- Structured data about data (DCMI – Dublin Core Metadata Initiative, 2003)
- Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO, 2004)

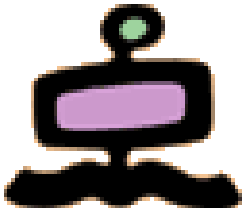
Resources and metadata



Metadata can be associated with any “entity”: physical, digital, abstract resource, etc.

- HTML documents
- digital images
- databases
- books
- museum objects
- archival records
- metadata records
- Web sites
- collections
- services
- physical places
- people
- institutions
- abstract “works”
- concepts
- events

What are metadata for (library and archive perspective)



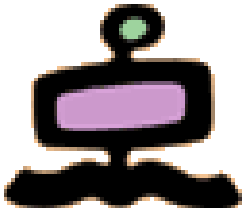
In relation to the ***resource*** (user tasks):

- to **find** entities that correspond to the user's stated search criteria (i.e., to locate either a single entity or a set of entities in a file or database as the result of a search using an attribute or relationship of the entity)
- to **identify** an entity (i.e., to confirm that the entity described corresponds to the entity sought, or to distinguish between two or more entities with similar characteristics)
- to **select** an entity that is appropriate to the user's needs (i.e., to choose an entity that meets the user's requirements with respect to content, physical format, etc., or to reject an entity as being inappropriate to the user's needs)
- to acquire or **obtain** access to the entity described (i.e., to acquire an entity through purchase, loan, etc., or to access an entity electronically through an online connection to a remote computer

In relation to the ***context***:

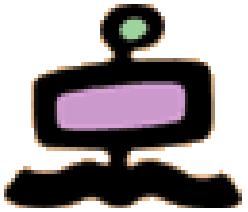
- to document the production, maintenance, distribution, archiving of the resource
- to provide its cultural, technical, administrative, structural background
- to preserve its integrity through medium and long term archiving
- to manage dates, agents, structures etc.
- to enable profiling, data mining etc.

Another point of view



"Then there is the question of cataloguing and metadata.
My view of the latter is that it is an ill-considered attempt to find some kind of Third Way between the wilderness of search engines and free text searching and the grand architecture of bibliographic control that librarians have developed over the last 150 years.
I think that metadata is the product of those with no knowledge of, or regard for, cataloguing; they are bibliographic alchemists seeking the philosopher's stone that will offer us effective cataloguing without expense and effective access without controlled vocabularies.
There is no such thing and the sooner that notion is disposed of, the better"
(Michael Gorman, past Dean of Library Services at California State University, past President of the American Library Association, November 2000)

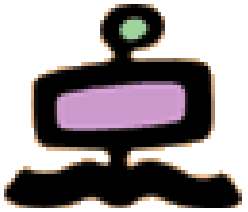
Not simply a cataloguing record



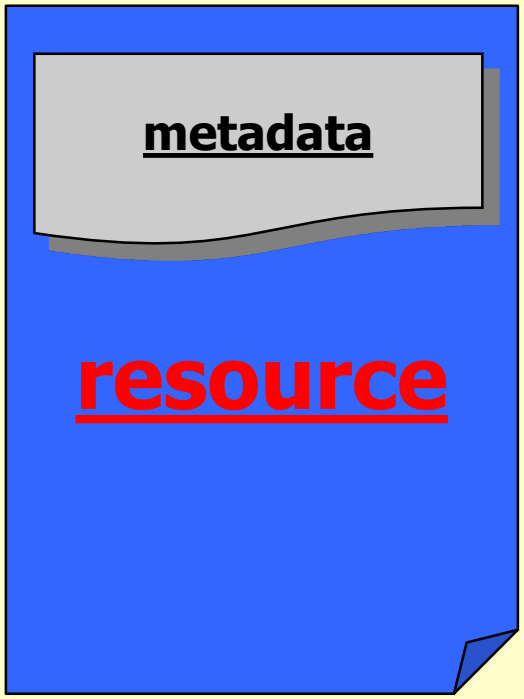
- An important reason for creating descriptive metadata is to facilitate discovery of relevant information, as it serves the same functions in resource discovery as good cataloguing does by:
  - allowing resources to be found by relevant criteria
  - identifying resources
  - bringing similar resources together
  - distinguishing dissimilar resources
  - giving location information

- In addition to resource discovery, metadata can
  - help organize electronic resources
  - facilitate interoperability and legacy resource integration
  - provide digital identification
  - support archiving and preservation

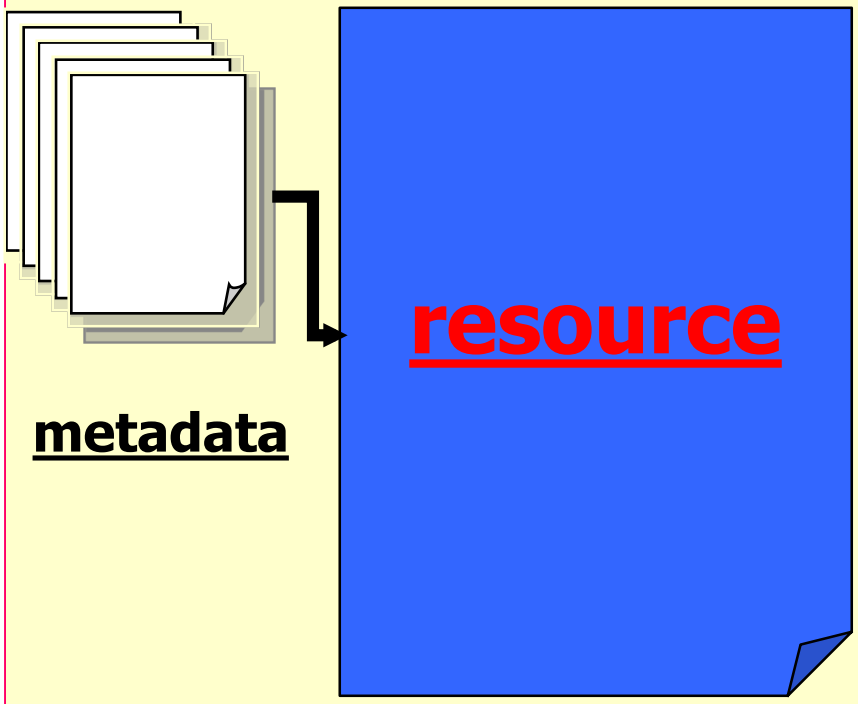
Storing Metadata



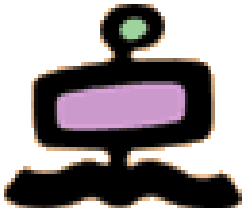
embedded metadata



stand-alone metadata

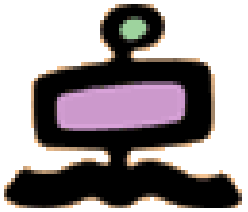


Storing metadata



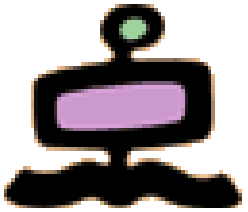
- Metadata can be embedded in a digital object or it can be stored separately. Metadata is often embedded in HTML documents and in the headers of image files
- Storing metadata with the object it describes ensures the metadata will not be lost, obviates problems of linking between data and metadata, and helps ensure that the metadata and object will be updated together
- However, it is impossible to embed metadata in some types of objects (for example, artifacts). Also, storing metadata separately can simplify the management of the metadata itself and facilitate search and retrieval. **Therefore, metadata is commonly stored in a database system and linked to the objects described**

Possible types of metadata



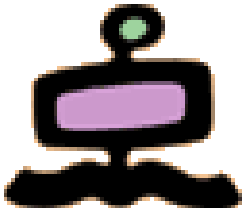
- *Descriptive metadata* describe a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords
- *Structural metadata* indicate how compound objects are put together, for example, how pages are ordered to form chapters
- *Technical metadata* indicate hardware or software used in converting an item/object to a digital format, or in storing or displaying it
- *Administrative metadata* provide information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. Two subsets of administrative data are:
  - *Rights management metadata,* which deal with intellectual property rights
  - *Preservation metadata,* which contain information needed to archive and preserve a resource

Dublin Core Metadata Initiative



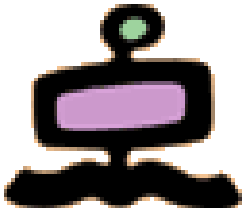
- Dublin Core
  - Dublin: Dublin, Ohio, 1995
  - Core: minimal set of broad and generic elements
- Dublin Core was originally developed with an eye to describing document-like objects (in institutional repositories)
  - Descriptions easy to create (unlike MARC)
- Despite initial focus, it has proved to be general enough to describe “any” type of objects
  - unlike catalog records, often tied to specific application fields
- It is now a widely used international standard
  - ISO Standard 15836-2003
  - NISO Standard Z39.85-2007
  - IETF RFC 5013 (2007)

IETF, RFC and W3C



- IETF: Internet Engineering Task Force
  - Groups of experts proposing and defining new technologies and applications in the web, that might become “Internet standards”
  - “We believe in running code and rough consensus”
- RFC: Request For Comments
  - The way to publish and define the standards in the web
  - Publications numbered sequentially starting with 1
    - RFC 1 in April 1969
    - RFC 9661 in September 2024
- W3C: World Wide Web Consortium
  - Consortium of industries and universities to ensure compatibility in the adoption of new standards
  - Issues recommendations (W3C standards) and certifications

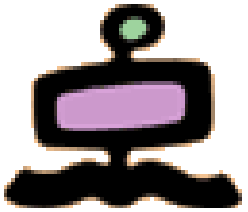
Fifteen Core Elements (1996)



Definition of elements (or terms) to describe resources

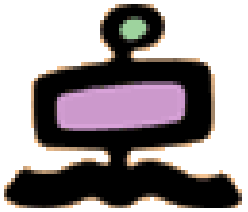
Content	**Intellectual Property**	**Instantiation**
Title Subject Description Type Source Relation Coverage	Creator Contributor Publisher Rights	Date Format Identifier Language

Characteristics of the Dublin Core



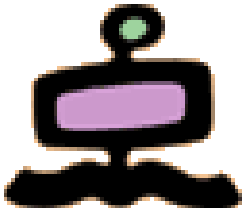
- All elements optional
- All elements repeatable
- Elements may be displayed in any order
- International in scope
- Extensible (Qualified Dublin Core)
- Dublin Core Principles
  - Dumb-Down
  - One-to-One
  - Appropriate Values

Simple and Qualified Dublin Core



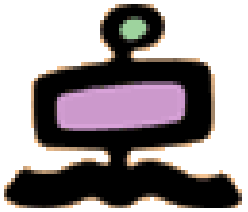
- Simple Dublin Core is limited to the original 15 elements
- Qualified Dublin Core includes, in addition:
  - New Elements
  - Qualifiers
    - Element Refinements
    - Encoding Schemes
      - Syntax Encoding Scheme
      - Vocabulary Encoding Scheme

New elements



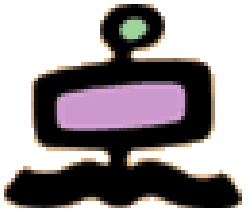
- Over time, the initial set of 15 elements has been enriched with a small number of (carefully selected) new elements
  - Audience
  - Provenance
  - RightsHolder
  - Instructional Method
  - Accrual Method
  - Accrual Periodicity
  - Accrual Policy

Qualifiers



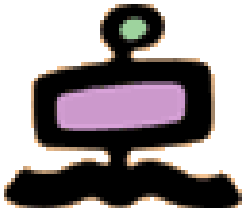
- Element refinements
  - These qualifiers make the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element, but with a more restricted scope.
- Encoding schemes
  - These qualifiers identify schemes that aid in the interpretation of an element value.
  - Syntax encoding schemes
    - A value expressed using a syntax encoding scheme will be a string formatted in accordance with a formal notation (e.g., "2000-01-01" as the standard expression of a date).
  - Vocabulary encoding schemes
    - A value expressed using a vocabulary encoding scheme will be a token selected from a controlled vocabulary (e.g., a term from a classification system or set of subject headings)
  - If an encoding scheme is not understood by a client or agent, the value may still be useful to a human reader

Element refinements



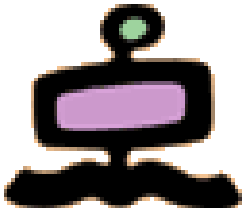
DCMES Element	Element Refinement(s)
Title	Alternative
Description	Table Of Contents Abstract
Date	Created Valid Available Issued Modified Date Accepted Date Copyrighted Date Submitted
Format	Extent Medium
Identifier	Bibliographic Citation
Relation	Is Version Of Has Version Is Replaced By Replaces Is Required By Requires Is Part Of Has Part Is Referenced By References Is Format Of Has Format Conforms To
Coverage	Spatial Temporal
Rights	AccessRights License
Audience	Mediator Education Level

Qualifiers



- Element refinements
  - These qualifiers make the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element, but with a more restricted scope.
- Encoding schemes
  - These qualifiers identify schemes that aid in the interpretation of an element value.
  - Syntax encoding schemes
    - A value expressed using a syntax encoding scheme will be a string formatted in accordance with a formal notation (e.g., "2000-01-01" as the standard expression of a date).
  - Vocabulary encoding schemes
    - A value expressed using a vocabulary encoding scheme will be a token selected from a controlled vocabulary (e.g., a term from a classification system or set of subject headings)
  - If an encoding scheme is not understood by a client or agent, the value may still be useful to a human reader

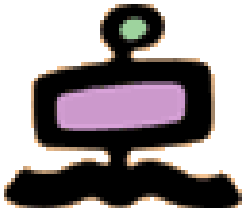
Encoding schemes



DCMES Element	Element Encoding Scheme(s)
Subject	LCSH MeSH DDC LCC UDC
Date	DCMIPeriod W3C-DTF
Type	DCMI Type Vocabulary
Format	IMT (MIME Types)
Identifier	URI
Source	URI

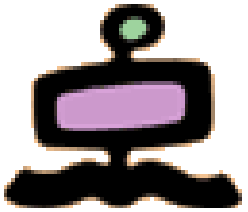
Language	ISO 639-2 RFC 3066
Relation	URI
Coverage spatial	DCMI Point ISO 3166 DCMI Box TGN
Coverage temporal	DCMI Period W3C-DTF
Rights license	URI

Characteristics of the Dublin Core



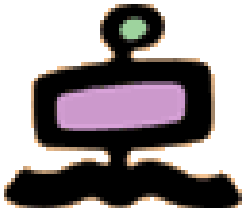
- All elements optional
- All elements repeatable
- Elements may be displayed in any order
- International in scope
- Extensible
  - Element refinements
  - Encoding schemes
- Dublin Core Principles
  - Dumb-Down
  - One-to-One
  - Appropriate Values

The Dumb-Down principle



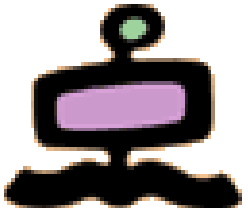
- The fifteen core elements are usable with or without qualifiers
  - Qualifiers make elements more specific:
  - Element Refinements narrow meanings, never extend
  - Encoding Schemes give context to element values
- If your software encounters an unfamiliar qualifier, look it up –or just ignore it!

The One-to-One principle



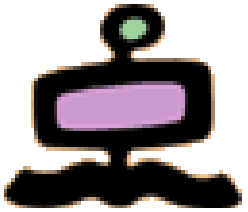
- Describe one manifestation of a resource with one record
  - Example: a digital image of the Mona Lisa is not described as if it were the same as the original painting
- Separate descriptions of resources from descriptions of the agents responsible for those resources
  - Example: email addresses and affiliations of creators are attributes of the creator, not the resource
- If needed, group related descriptions into a “description set” (record)

The “Appropriate Values” priciple



- Use elements, element refinements and qualifiers to meet the needs of your local context, but. . .
- Remember that your metadata may be interpreted by machines and people, so. . .
- Consider whether the values you use will aid discovery outside your local context and. . .
- Make decisions about your local practices accordingly

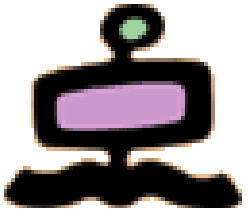
Fifteen Core Elements (1996)



Definition of elements (or terms) to describe resources

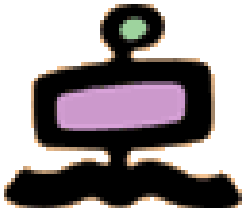
Content	**Intellectual Property**	**Instantiation**
Title	Creator	Date
Subject	Contributor	Format
Description	Publisher	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

Element “Title”



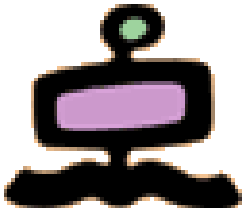
- Description:
  - The name given to the resource. Typically, a Title will be a name by which the resource is formally known.
- Guidelines for creation of content:
  - If in doubt about what constitutes the title, repeat the Title element and include the variants in second and subsequent Title iterations.
  - If the item is in HTML, view the source document and make sure that the title identified in the title header (if any) is also included as a Title.
- Examples:
  - Title="A Pilot's Guide to Aircraft Insurance"
  - Title="The Sound of Music"
  - Title="Green on Greens"
  - Title="AOPA's Tips on Buying Used Aircraft”

Element “Subject”



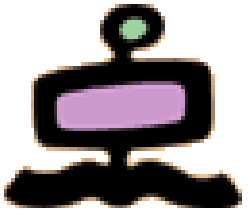
- Description
  - The topic of the content of the resource. Typically, a Subject will be expressed as keywords or key phrases or classification codes that describe the topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
- Guidelines for creation of content
  - Select subject keywords from the Title or Description information, or from within a text resource. If the subject of the item is a person or an organization, use the same form of the name as you would if the person or organization were a Creator or Contributor.
  - In general, choose the most significant and unique words for keywords, avoiding those too general to describe a particular item. Subject might include classification data if it is available (for example, Library of Congress Classification Numbers or Dewey Decimal numbers) or controlled vocabularies (such as Medical Subject Headings or Art and Architecture Thesaurus descriptors) as well as keywords.
  - When including terms from multiple vocabularies, use separate element iterations. If multiple vocabulary terms or keywords are used, either separate terms with semi-colons or use separate iterations of the Subject element.
- Examples
  - Subject="Aircraft leasing and renting"
  - Subject="Dogs"
  - Subject="Olympic skiing"
  - Subject="Street, Picabo"

Element “Description”



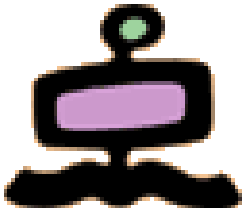
- **Description**
  - **An account of the content of the resource. Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.**
- **Guidelines for creation of content**
  - **Since the Description field is a potentially rich source of indexable terms, care should be taken to provide this element when possible. Best practice recommendation for this element is to use full sentences, as description is often used to present information to users to assist in their selection of appropriate resources from a set of search results.**
  - **Descriptive information can be copied or automatically extracted from the item if there is no abstract or other structured description available. Although the source of the description may be a web page or other structured text with presentation tags, it is generally not good practice to include HTML or other structural tags within the Description element. Applications vary considerably in their ability to interpret such tags, and their inclusion may negatively affect the interoperability of the metadata.**
- **Examples:**
  - **Description="Illustrated guide to airport markings and lighting signals, with particular reference to SMGCS (Surface Movement Guidance and Control System) for airports with low visibility conditions.“**
  - **Description="Teachers Domain is a multimedia library for K-12 science educators, developed by WGBH through funding from the National Science Foundation as part of its National Science Digital Library initiative. The site offers a wealth of classroom-ready instructional resources, as well as online professional development materials and a set of tools which allows teachers to manage, annotate, and share the materials they use in classroom teaching."**

Element “Creator”



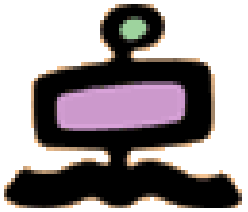
- **Description**
  - **An entity primarily responsible for making the content of the resource. Examples of a Creator include a person, an organization, or a service. Typically the name of the Creator should be used to indicate the entity.**
- **Guidelines for creation of content:**
  - **Creators should be listed separately, preferably in the same order that they appear in the publication. Personal names should be listed surname or family name first, followed by forename or given name. When in doubt, give the name as it appears, and do not invert.**
  - **In the case of organizations where there is clearly a hierarchy present, list the parts of the hierarchy from largest to smallest, separated by full stops and a space. If it is not clear whether there is a hierarchy present, or unclear which is the larger or smaller portion of the body, give the name as it appears in the item.**
  - **If the Creator and Publisher are the same, do not repeat the name in the Publisher area. If the nature of the responsibility is ambiguous, the recommended practice is to use Publisher for organizations, and Creator for individuals. In cases of lesser or ambiguous responsibility, other than creation, use Contributor.**
- **Examples:**
  - **Creator="Shakespeare, William"**
  - **Creator="Wen Lee"**
  - **Creator="Hubble Telescope"**
  - **Creator="Internal Revenue Service. Customer Complaints Unit"**

Element “Identifier”



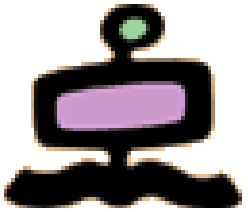
- Description
  - An unambiguous reference to the resource within a given context. Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system.
  - Examples of formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).
- Guidelines for content creation
  - This element can also be used for local identifiers (e.g. ID numbers or call numbers) assigned by the Creator of the resource to apply to a particular item. It should not be used for identification of the metadata record itself.
- Examples:
  - Identifier="http://purl.oclc.org/metadata/dublin_core/& quot;
  - Identifier="ISBN:0385424728“
  - Identifier="H-A-X 5690B" [publisher number]

Element “Type”



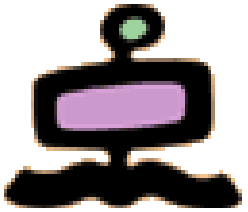
- Description
  - The nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMI Type vocabulary). To describe the physical or digital manifestation of the resource, use the FORMAT element.
- Guidelines for content creation:
  - If the resource is composed of multiple mixed types then multiple or repeated Type elements should be used to describe the main components.
  - Because different communities or domains are expected to use a variety of type vocabularies, best practice to ensure interoperability is to include at least one general type term from the DCMI Type vocabulary in addition to the domain specific type term(s), in separate Type element iterations.

DCMI Types



- Collection
- Dataset
- Event
- Image
- InteractiveResource
- MovingImage
- PhysicalObject
- Service
- Software
- Sound
- StillImage
- Text

Examples of element “Type”

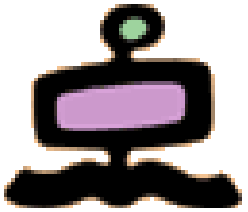


- Example 1 (a simulator program, e.g. a game)
  - Type="Image“ (DCMI Type Vocabulary)
  - Type="Sound“ (DCMI Type Vocabulary)
  - Type="Text“(DCMI Type Vocabulary)
  - Type="simulation“ (unspecified source)

- Example 2 (an Electronic art exhibition catalog)
  - Type="Image“ (DCMI Type Vocabulary)
  - Type="Text“ (DCMI Type Vocabulary)
  - Type="Exhibition catalog" (unspecified source)

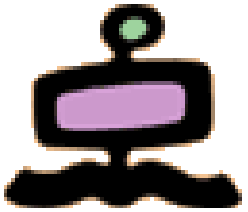
- Example 3 (a Multimedia educational program with interactive assignments)
  - Type="Image“ (DCMI Type Vocabulary)
  - Type="Text“(DCMI Type Vocabulary)
  - Type="Software“(DCMI Type Vocabulary)
  - Type="InteractiveResource“(DCMI Type Vocabulary)

Element “Format”



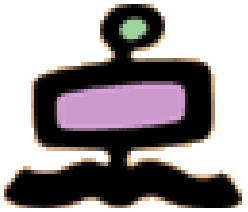
- Description
  - The physical or digital manifestation of the resource. Typically, Format may include the media-type or dimensions of the resource. Examples of dimensions include size and duration. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource.
  - Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types defining computer media formats).
- Guidelines for content creation
  - In addition to the specific physical or electronic media format, information concerning the size of a resource may be included in the content of the Format element if available. In resource discovery, size, extent or medium of the resource might be used as a criterion to select resources of interest, since a user may need to evaluate whether they can make use of the resource within the infrastructure available to them.
  - When more than one category of format information is included in a single record, they should go in separate iterations of the element.

Examples of element “Format”



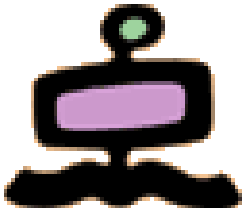
- Example1:
  - Title="Dublin Core icon"
  - Identifier="http://purl.org/metadata/dublin_core/images/dc2.gif "
  - Type="Image"
  - Format="image/gif"
  - Format="4 KB"
- Example 2
  - Subject="Saturn"
  - Type="Image"
  - Format="image/gif "
  - Format="40 x 512 pixels"
  - Identifier="http://www.not.iac.es/newwww/photos/images/satnot.gif "
- Example 3
  - Title="The Bronco Buster"
  - Creator="Frederic Remington"
  - Type="Physical object"
  - Format="bronze"
  - Format="22 in."

Element “Coverage”



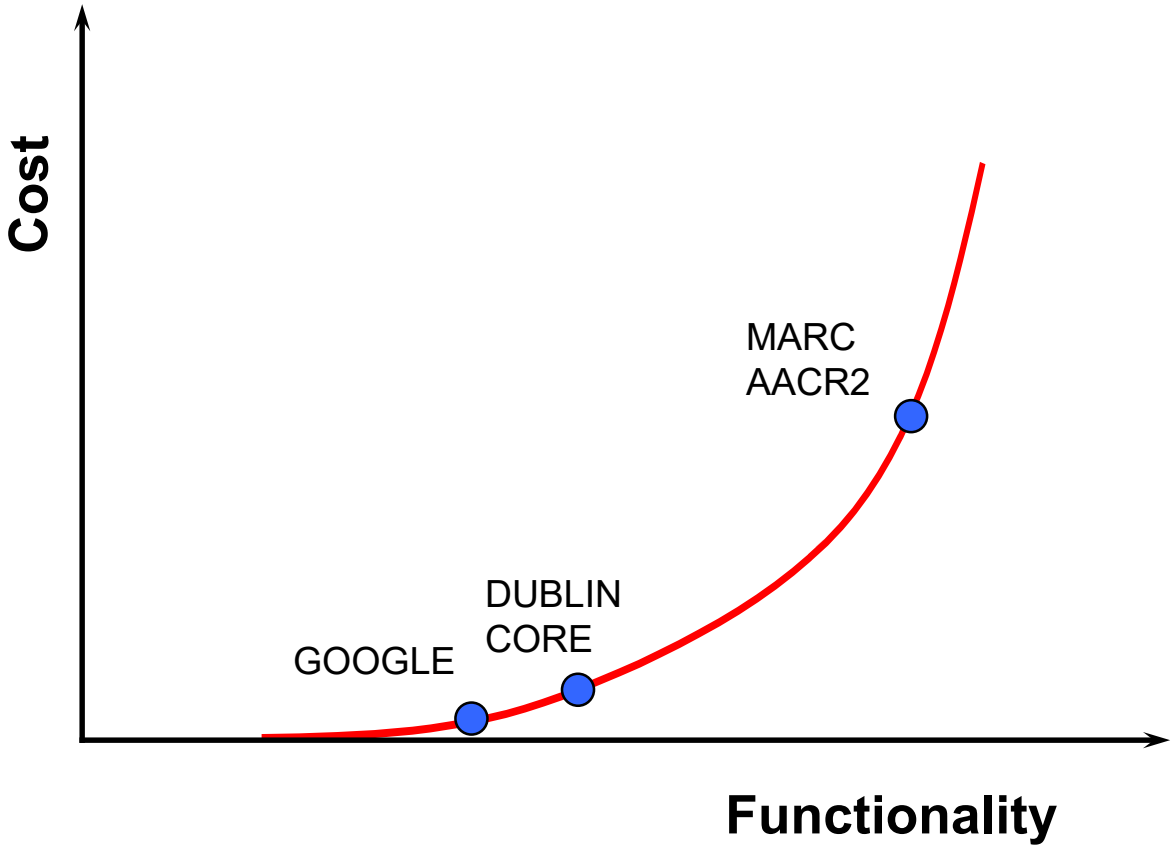
- Description
  - The extent or scope of the content of the resource. Coverage will typically include spatial location (a place name or geographic co-ordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity).
  - Recommended best practice is to select a value from a controlled vocabulary (for example, the Getty Thesaurus of Geographic Names.
  - Where appropriate, named places or time periods should be used in preference to numeric identifiers such as sets of co-ordinates or date ranges.
- Guidelines for content creation:
  - Whether this element is used for spatial or temporal information, care should be taken to provide consistent information that can be interpreted by human users, particularly in order to provide interoperability in situations where sophisticated geographic or time-specific searching is not supported.
  - For most simple applications, place names or coverage dates might be most useful. For more complex applications, consideration should be given to using an encoding scheme that supports appropriate specification of information, such as DCMI Period, DCMI Box or DCMI Point.
- Examples:
  - Coverage="1995-1996"
  - Coverage="Boston, MA"
  - Coverage="17th century"
  - Coverage="Upstate New York"

MARC and Dublin Core

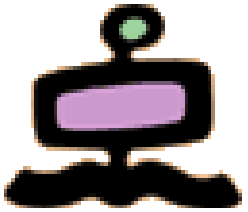


- They represent opposite ends of the complexity spectrum.
- MARC is a comprehensive, well-developed, carefully controlled scheme intended to be generated by professional catalogers for use in libraries.
- Dublin Core is an intentionally minimalist standard intended to be applied to a wide range of digital library materials by people who are not trained in library cataloging.
- These two schemes are of interest not only for their practical value, but also to highlight diametrically opposed underlying philosophies.

Defining metadata

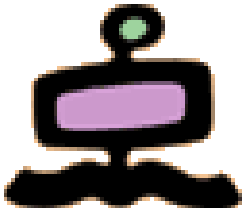


Expressing Dublin Core



- Expressing Dublin Core in HTML/XHTML meta and link elements
  - Status: DCMI Recommendation
  - Date Issued: 2008-08-04
  - Identifier: https://www.dublincore.org/specifications/dublin-core/dc-html/
- Expressing Dublin Core using XML
  - Status: Proposed Recommendation
  - Date Issued: 2008-09-01
  - Identifier: https://www.dublincore.org/specifications/dublin-core/dc-ds-xml/
- Expressing Dublin Core using RDF (we will see later)
  - Status: DCMI Recommendation
  - Date Issued: 2008-01-14
  - Identifier: https://www.dublincore.org/specifications/dublin-core/dc-rdf/

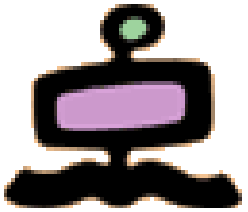
Metadata in HTML



```
<head>
  <meta charset="UTF-8">
  <meta name="title" content="Example">
  <meta name="creator" content="Joe Smith">
  <meta name="description" content="blah blah">
</head>

<body>
    ..........
</body>
```

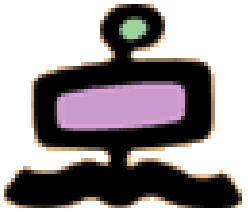
Simple Dublin Core in XML



```
<?xml version="1.0"?>
<metadata
  xmlns ="http://example.org/myapp/"
  xmlns:xsi ="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation ="http://example.org/myapp/
                       http://example.org/myapp/schema.xsd"
  xmlns:dc ="http://purl.org/dc/elements/1.1/">
  <dc:title>
    UKOLN
  </dc:title>
  <dc:description>
    UKOLN is a national focus of expertise ......
  </dc:description>
  <dc:publisher>
    UKOLN, University of Bath
  </dc:publisher>
  <dc:identifier>
    http://www.ukoln.ac.uk/
  </dc:identifier>
</metadata>
```

The resource being described is the home page of UKOLN

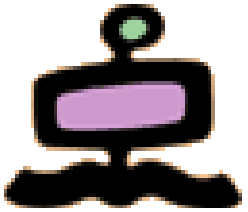
Qualified Dublin Core in XML



```
<?xml version="1.0"?>
<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://example.org/myapp/ http://example.org/myapp/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/">
  <dc:title>UKOLN</dc:title>
  <dcterms:alternative>
    UK Office for Library and Information Networking</dcterms:alternative>
  <dc:subject>national centre, network information support, etc. </dc:subject>
  <dc:subject xsi:type="dcterms:DDC">062</dc:subject>
  <dc:subject xsi:type="dcterms:UDC">061(410)</dc:subject>
  <dc:description>UKOLN is a national focus .....</dc:description>
  <dc:description xsi:lang="fr">
    UKOLN est un centre national d'expertise .....</dc:description>
  <dc:publisher>UKOLN, University of Bath</dc:publisher>
  <dcterms:isPartOf xsi:type="dcterms:URI">
    http://www.bath.ac.uk/</dcterms:isPartOf>
  <dc:identifier xsi:type="dcterms:URI">
    http://www.ukoln.ac.uk/</dc:identifier>
  <dcterms:modified xsi:type="dcterms:W3CDTF">2001-07-18</dcterms:modified>
  <dc:format xsi:type="dcterms:IMT">text/html</dc:format>
  <dcterms:extent>14 Kbytes</dcterms:extent>
</metadata>
```

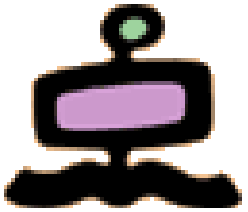
The resource being described is the home page of UKOLN

Dublin Core User Guide(s)



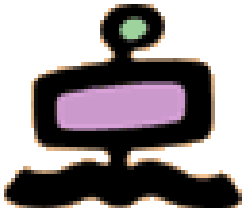
- Dublin Core User Guide 2005
  - First user guide, simple and easy to read
  - https://www.dublincore.org/specifications/dublin-core/usageguide/
- Dublin Core User Guide 2011
  - First major revision of User Guide, introducing RDF and Linked data
- Dublin Core User Guide 2019
  - Current version, describing the use of Dublin Core for creating metadata and the use of Dublin Core for publishing metadata
  - https://www.dublincore.org/resources/userguide/

Where are we



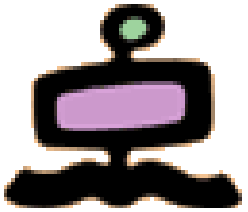
- Digital Libraries
  - Access to information
    - Describing Information
      - Catalog records (libraries)
        - MARC
      - Metadata (the Web)
        - Dublin Core
        - MODS
        - METS
        - EAD
        - TEI
        - .....

Metadata standards for DLs



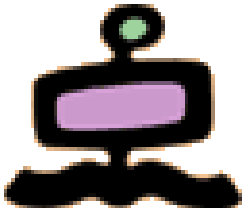
- Interoperability and object exchange requires the use of established standards
- Many digital objects are complex and are comprised of multiple files
- Complex digital objects require more sets of metadata than physical objects for their management and use
  - Descriptive
  - Technical
  - Structural
  - Administrative
    - Digital provenance/events
    - Rights/Terms and conditions
- XML is (was) the de-facto standard for metadata descriptions on the Internet

XML Resource Description Formats (Library of Congress)



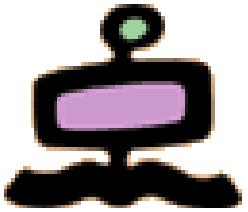
- MARCXML
  - XML representation of MARC 21 data
- MODS (Metadata Object Description Standard)
  - XML markup for metadata from existing MARC 21 records and original resource description
- MADS (Metadata Authority Description Standard)
  - XML markup for authority data from MARC 21 records and original authority data
- EAD (Encoded Archival Description)
  - XML markup designed for encoding archival finding aids

MARCXML



- Millions of rich descriptive records in MARC systems: can be reused in an XML environment using MARCXML
- MARCXML uses the MARC data element set in an XML syntax
- Lossless roundtrip conversion to/from MARC 21 record
- Allows interoperability with other XML schemes by taking advantage of free XML tools
- Allows for collaborative use of metadata for access (e.g. OAI)

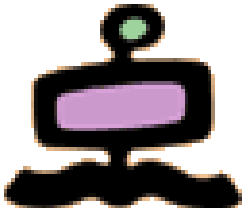
MODS



MODS - Metadata Object Description Schema

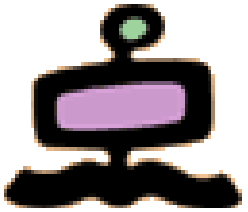
- An XML descriptive metadata standard
- Can be viewed as a (simple) alternative to the MARC format
  - Contains a subset of MARC data elements
- MODS does not assume the use of any specific rules for description
- Element set is particularly applicable to digital resources
- Note to catalogers:
  "MODS does not make you obsolete! The same knowledge and skills needed for traditional cataloging (AACR, LCSH, controlled vocabularies, etc.) still apply. You will only need to learn a different syntax (i.e., different from MARC) for expressing bibliographic information in machine-readable form."

MODS high-level elements



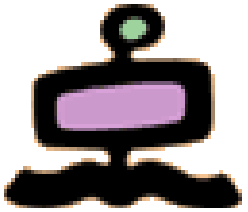
- Title Info
- Name
- Type of resource
- Genre
- Origin Info
- Language
- Physical description
- Abstract
- Table of contents
- Target audience
- Note
- Subject
- Classification
- Related item
- Identifier
- Location
- Access conditions
- Part
- Extension
- Record Info

EAD
Encoded Archival Description



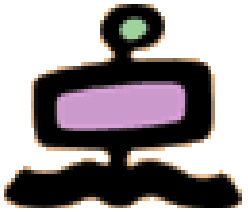
Abbreviation -- <abbr> Abstract -- <abstract> Accruals -- <accruals> Acquisition Information -- <acqinfo> Address -- <address> Address Line -- <addressline> Alternative Form Available -- <altformavail> Appraisal Information -- <appraisal> Arc -- <arc> Archival Description -- <archdesc> Archival Description Group -- <archdescgrp> Archival Reference -- <archref> Arrangement -- <arrangement> Author -- <author> Bibliographic Reference -- <bibref> Bibliographic Series -- <bibseries> Bibliography -- <bibliography> Biography or History -- <bioghist> Block Quote -- <blockquote> Change -- <change> Chronology List -- <chronlist> Chronology List Item -- <chronitem> Component (Eighth Level) -- <c08> Component (Eleventh Level) -- <c11> Component (Fifth Level) -- <c05> Component (First Level) -- <c01> Component (Fourth Level) -- <c04> Component (Ninth Level) -- <c09> Component (Second Level) -- <c02> Component (Seventh Level) -- <c07> Component (Sixth Level) -- <c06> Component (Tenth Level) -- <c10> Component (Third Level) -- <c03> Component (Twelfth Level) -- <c12> Component (Unnumbered) -- <c> Conditions Governing Access -- <accessrestrict> Conditions Governing Use -- <userestrict> Container -- <container> Controlled Access Headings -- <controlaccess> Corporate Name -- <corpname> Creation -- <creation> Custodial History -- <custodhist> Date -- <date> Date of the Unit -- <unitdate> Definition List Item -- <defitem> Description Group -- <descgrp> Description of Subordinate Components -- <dsc> Description of Subordinate Components Group -- <dscgrp> Descriptive Identification -- <did> Descriptive Rules -- <descrules> Digital Archival Object -- <dao> Digital Archival Object Description -- <daodesc> Digital Archival Object Group -- <daogrp> Digital Archival Object Location -- <daoloc> Dimensions -- <dimensions> EAD Group -- <eadgrp> EAD Header -- <eadheader> EAD Identifier -- <eadid> Edition -- <edition> Edition Statement -- <editionstmt> Emphasis -- <emph> Encoded Archival Description -- <ead> Event -- <event> Event Group -- <eventgrp> Expansion -- <expan> Extended Pointer -- <extptr> Extended Pointer Location -- <extptrloc> Extended Reference -- <extref> Extended Reference Location -- <extrefloc> Extent -- <extent> Family Name -- <famname> File Description -- <filedesc> File Plan -- <fileplan> First Heading -- <head01> Front Matter -- <frontmatter> Function -- <function> Genre/Physical Characteristic -- <genreform> Geographic Name -- <geogname> Heading -- <head> ID of the Unit -- <unitid> Imprint -- <imprint> Index -- <index> Index Entry -- <indexentry> Item -- <item> Label -- <label> Language -- <language> Language of the Material -- <langmaterial> Language Usage -- <langusage> Legal Status -- <legalstatus> Line Break -- <lb> Linking Group -- <linkgrp> List -- <list> List Heading -- <listhead> Location of Originals -- <originalsloc> Material Specific Details -- <materialspec> Name -- <name> Name Group -- <namegrp> Note -- <note> Note Statement -- <notestmt> Number -- <num> Occupation -- <occupation> Origination -- <origination> Other Descriptive Data -- <odd> Other Finding Aid -- <otherfindaid> Paragraph -- <p> Personal Name -- <persname> Physical Characteristics and Technical Requirements -- <phystech> Physical Description -- <physdesc> Physical Facet -- <physfacet> Physical Location -- <physloc> Pointer -- <ptr> Pointer Group -- <ptrgrp> Pointer Location -- <ptrloc> Preferred Citation -- <prefercite> Processing Information -- <processinfo> Profile Description -- <profiledesc> Publication Statement -- <publicationstmt> Publisher -- <publisher> Reference -- <ref> Reference Location -- <refloc> Related Material -- <relatedmaterial> Repository -- <repository> Resource -- <resource> Revision Description -- <revisiondesc> Runner -- <runner> Scope and Content -- <scopecontent> Second Heading -- <head02> Separated Material -- <separatedmaterial> Series Statement -- <seriesstmt> Sponsor -- <sponsor> Subject -- <subject> Subordinate Area -- <subarea> Subtitle -- <subtitle> Table -- <table> Table Body -- <tbody> Table Column Specification -- <colspec> Table Entry -- <entry> Table Group -- <tgroup> Table Head -- <thead> Table Row -- <row> Text Division -- Title -- <title> Title of the Unit -- <unittitle> Title Page -- <titlepage> Title Proper of the Finding Aid -- <titleproper> Title Statement -- <titlestmt>

XML Digital Library Standards (Library of Congress)



- METS (Metadata Encoding & Transmission Standard)
  - Structure for encoding descriptive, administrative, and structural metadata
- MIX (NISO Metadata for Images in XML)
  - XML schema for encoding technical data elements required to manage digital image collections
- PREMIS (Preservation Metadata)
  - Data dictionary and supporting XML schemas for core preservation metadata needed to support the long-term preservation of digital materials.
- TextMD (Technical Metadata for Text)
  - XML schema that details technical metadata for text-based digital objects
- AudioMD and VideoMD
  - XML schemas for technical metadata on audio- and video-based digital objects

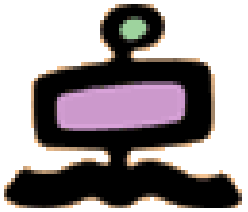
METS



Metadata Encoding and Transmission Standard

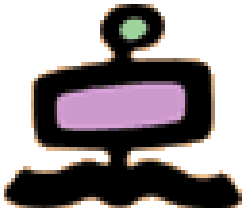
- An XML Schema for the purpose of creating XML document instances that express
  - the hierarchical structure of digital library objects
  - the names and locations of the files that comprise the digital object
  - the associated metadata (e.g., MODS)
- METS can be used as a tool for modeling real world objects, such as specific document types

Structure of a METS document



- There are 7 sections in a METS document
  - <mets>
    - <metsHdr/> - METS header (document talks about itself)
    - <dmdSec/> - Descriptive metadata (MARCXML, MODS, etc.)
    - <amdSec/> - Administrative metadata (copyright info., etc.)
    - <fileSec/> - File section (names and locations of files)
    - <structMap/> - Structural map (relationships of the parts)
    - <structLink/> - Linking information
    - <behaviorSec/> - Binding executables/actions to object
  - </mets>

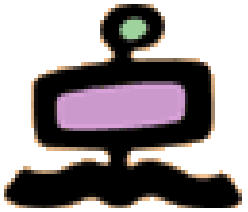
MIX



MIX – Metadata for Images in XML

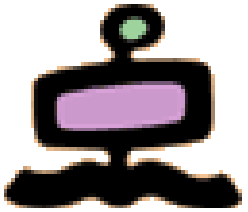
- An XML Schema designed for expressing technical metadata for digital still images (pictures)
- Based on the NISO Z39.87 Data Dictionary – Technical Metadata for Digital Still Images
- Used to express attributes of digital images such as file format, file size, dimensions, resolution, compression, etc.
- Can be used standalone or as an extension schema with METS

MIX main elements



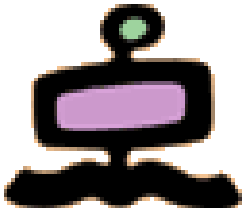
- Basic Digital Object Information
  - ObjectIdentifier
  - fileSize
  - FormatDesignation
  - FormatRegistry
  - byteOrder
  - Compression
  - Fixity
- Basic Image Information
  - BasicImageCharacteristics
  - SpecialFormatCharacteristics
- Image Capture Metadata
  - SourceInformation
  - GeneralCaptureInformation
  - ScannerCapture
  - DigitalCameraCapture
  - orientation
  - methodology
- Image Assessment Metadata
  - SpatialMetrics
  - ImageColorEncoding
  - TargetData
- Change History
  - ImageProcessing
  - PreviousImageMetadata

Where we came from



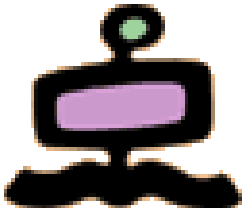
- Classification and Cataloguing
  - Classification means to bring related items together. Conventional libraries, in order to stack books on related subjects together, have used library classification. This facilitates the browsing approach of the information seekers
  - Cataloguing creates “document surrogates”, i.e. a description of a document (a catalog record), to be used (to a certain extent) in the place of the document. Catalog records provide searching facility by Authors, Titles, Subjects, Series and other elements. Classification use notation symbols, and Authority Files provide recommended vocabularies

Where we have arrived



- Digital Libraries
  - Access to information
    - Describing Information
      - Catalog records (libraries)
        - MARC
      - Metadata (the Web)
        - Dublin Core
        - MODS
        - METS
        - EAD
        - TEI
        - ……

What we have learned



- Many types of Metadata
  - Administrative
  - Descriptive
  - Access/Use
  - Preservation
  - Technical/Structural
  - Other ...
- Many metadata schema in the world of Digital Libraries
  - Dublin Core
  - MARCXML
  - MODS
  - TEI
  - EAD
  - Etc.
- Most used representation (expression) of metadata (unitl recently)
  - XML (moving to JSON and to RDF)