Unit 1: Resource Description for OA Resources

Objective: This unit is meant for helping you in application of standard metadata schemas in organizing OA resources.

Learning Outcomes

After going through this topic, the participants are expected to be able to:

- Define metadata;
- Identify and describe the elements of some important metadata description formats;
- Understand policies related to metadata applications;
- Critically examine the scopes of generic and domain-specific metadata schemas for organizing OA resources;
- Explain the roles of models, crosswalks and interoperability standards in metadata applications including the scope of emerging initiatives in OA metadata landscape;
- Explore the software-level application of metadata in organizing OA resources.

Areas of focus

- Policies Resource Description
- Open Access and Metadata
    - Policy Framework
    - Application Framework
    - Usage Metadata
- Generic Metadata Schema
- Domain-specific Metadata Schemas
    - Learning Objects Domain
    - Theses and Dissertations
    - Other Domains
- Metadata Modelling
    - Bibliographic Data Models
    - Applications of RDF and XML
- Application of Metadata in Open Access
    - Guidelines and Initiatives
    - Software-level applications
    - Authority Control in Gold OA and Green OA
- Metadata: Crosswalks and Interoperability Standards

Resource Description

- Metadata, in general, is referred to as data about data, and provides basic information such as the author of a work, the date of creation, links to any related works, etc.
- Metadata exists for almost every conceivable object or group of objects, whether stored in electronic form or not.
- In the library world, one easily identifiable form of metadata is the card catalogue; the information on the card is metadata about a book.
- In a traditional library, where cataloguing is the work of trained professionals, complex metadata schemes such as MARC, CCF etc. are used for description of library resources..”

Metadata defined

IFLA defines metadata as “The term refers to any data used to aid the identification, description and location of networked electronic resources”(IFLA, 2002). According to W3C “Metadata is the machine understandable information for the Web” (W3C, 2003).

Data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation” (Getty Research Institute; Murtha, 2002).

Typology of metadata

a) Descriptive metadata is meant to serve the purposes of discovery (i.e. how one can find a resource), identification (i.e. how a resource can be distinguished from other similar resources), selection (i.e. how to determine that a resource fills a particular need), collocation (bringing together related works), obtain (obtaining a copy of resource, or access to one) and other related functions (evaluation, linkage and usability).

b) Administrative metadata is information intended to facilitate the management of resources such as date of creation, rights and restrictions of access and archiving, control or processing activities etc.

c) Structural metadata is concerned with recording of relationships that holds compound digital objects together.

Metadata Schemas

Metadata schemas are set of metadata elements and rules for their use that have been defined for a particular purpose. A metadata schema specifies three independent but related aspects of metadata – semantics, content rules and syntax.

- **Semantics** refers to the metadata elements that are included in the schema by giving each of them a name and definition. A metadata schema also specifies whether each element is mandatory, optional or conditionally required and whether the element may or may not be repeated.
- **Content rules** indicate how values for metadata elements are selected and represented. For example, semantics of a metadata schema may define the element “author” but the content rules would specify which agents qualify as author (selection) and how an author’s name should be recorded (representation).

- **Syntax of a metadata schema** is concerned with the encoding of metadata elements in machine-readable form. Syntax also specifies the way of transmission, transport and communication of metadata between different systems.

Types of Metadata Schemas

a) **Generic** - are intended to be generally applicable to all types of resources (e.g., Dublin Core Metadata Elements Set) – e.g. Dublin Core Metadata Element set (DCMES)/ Dublin Core Metadata Initiative (DCMI) pp. 17 – 18)

Characteristics Generic metadata schemas

a. flexible enough for the description of variety of resources in different subject areas.

b. the meanings of the elements will be understood by most users.

This quality has been achieved by DC metadata by following Six Principles:

1. **Intrinsicality:** DC metadata is based on intrinsic data. These data refers to the property that could be identified from the intellectual content and physical form of the resource;
2. **Extensibility:** It allows inclusion of extra descriptive materials for specialized requirements;
3. **Syntax Independence**: It is applicable to a wide range of disciplines and application program;
4. **Optionality**: All the DC elements are optional;
5. **Repeatability:** All the DC elements are repeatable. For example, a resource with multiple authorship may use the "Creator" element repeatedly to accommodate all the authors; and
6. **Modifiability:** Each element in the Dublin Core has a definition, which is self explanatory. Each element can be modified by an optional qualifier and in such cases the definition of the element is modified by the value of the qualifier.

DCMES Major elements – pp. 19 - 23

Domain-specific - domain-specific metadata schemas are primarily designed to describe items related to a particular category (e.g. VRA [Visual Resource Association] Core for visual resource collection, FGDC (Federal Geographic Data Committee) metadata schema for geospatial data etc.).

Examples of DSs

- TEI (Text Encoding Initiative),
- FGDC (Federal Geographic Data Committee),
- GILS (Global Information Locator Service), OAI (Open Archive Initiative) etc.

Characteristics of Domain-Specific

- have been developed to operate within a narrowly defined subject field
- generally not suitable for the description of a wider range of resources.
- complex in nature and thereby geared towards creation by experts and interpretation by computers.

Nature of contents for different categories of metadata elements in schemas

- Descriptive metadata elements

- Administrative metadata elements,
- Structural metadata elements (Semantics)
- Content rules for metadata representation and
- Syntax for machine-readable metadata encoding.

Descriptive metadata elements

- Bibliographic description (such as Dublin Core, MODS, MARC21, MARCXML, ONIX schemas for metadata representation);
- Content description (such as DDI, SDMX, FGDC, EAD, TEI etc.);
- Description of structure, context and source of the data; information about the methods, instruments, and techniques used in the creation or collection of the data;
- References and links to publications pertaining to the data; and
- Information on how the data have been processed prior to submission to the repository.

Administrative metadata elements

- Preservation metadata to represent lifecycle of the data, recording of events related to submission, curation and dissemination (such as PREMIS) and event history data (for linking with digital objects) ;
- Rights management metadata;
- Technical metadata (storage format etc.); and
- Representation Information (internal coding, rendering data etc).

Structural metadata elements

- Structural metadata indicates relationships amongst different components of a set of associated data that are particularly important for Web aggregation. These aggregations are also called compound digital objects.
- FOXML (Standard in use for Fedora repository software, where compound objects are treated as a single file);
- OAI-ORE (An OAI initiative that defines compound objects distributed on the Internet through the creation of resource maps which use unique URLsfor each component; It has four basic components i) Resource (an item of interest); ii) URI (a global resource identifier); iii) Representation (a DataStream accessible through URI by using a protocol like HTTP); and iv) Link (a connection between two resources);
- METS (An LoC standard that is used as a 'wrapper' for compound digital objects and very useful for import/export in repositories); and
- RDF (A W3C standard that provides a simple way to represent Web resources, in the form of subject-predicate-object expressions that relate objects to one another).

Activity: *In groups discuss why understanding metadata as a librarian is important for Open access*

Policy Framework

Issues that need to be discussed in the policy framework for metadata

i) Who can enter or edit metadata?

ii) Which metadata standards are to be followed?

iii) Whether different metadata schemas are required for describing different type of documents?

iv) Whether or not the repository systems allow metadata harvesting by service providers?

v) Which protocols should OA system support for metadata harvesting?

Ref: pp. 10 – 12

Activity: *Study the Metadata policies in OARs – what is your opinion of the repository to the respective metadata policies pp. 13 - 15?*

Answer – Application framework – pp. 15 – 16

Theses and Dissertations Metadata Schemas

- ETD-MS - ETDMS is based on the Dublin Core Element Set, but includes an additional element specific to metadata regarding theses and dissertations - deals with both paper and electronic theses
- UK-ETD - is recommended by Electronic Theses Online Service (EThOS), UK. EThOS is the Electronic Theses Online System which allows individuals to find access and archive doctoral e-theses that are produced in UK Higher Education institutions
- Shodhganga - mainly used in Indian universities - to provide online accessibility to Indian theses for archiving and free access. The Shodhganga metadata schema has been developed as domainspecific schema to deal with ETDs of Indian universities.

Notes:

1) Comparison of three metadata standards for Theses and dissertations – pp. 29 – 30
2) Illustrative list of popular domain-specific metadata schemas – pp. 31 – 35

Activity: Study the metadata schemas on pp 31 – 35 and group them based on this domains - ETD, Image, Maps, Learning Objects, Cultural objects, and Compound digital objects.

Application of metadata in OA

- Most of the guidelines (as developed in US and UK) advocate to categorize metadata elements into four categories -
  - Required,
  - Required if Applicable,
  - Recommended and Optional.
- The basic purpose of the categorization is to identify the elements necessary for a user in a shared metadata environment.

- Guidelines are not format-specific; rather they identify those elements commonly needed across all formats.

An analysis of existing suggestions and guidelines shows the following categorization of metadata elements -

Required

- Date Created or Date Published (dc:date)
- Identifier (dc:identifier)
- Institution Name (dc:publisher)
- Title (dc:title)
- Type of Resource (dc:type)

Required if Applicable

- Creator (dc:creator)
- Extent (dc:format)
- Language of Resource (dc:language)
- Related Item (dc:relation)

Recommended

- Description (dc:description)
- Access or Use Restrictions (dc:rights)
- Format of Resource (dc:format)
- Place of Origin (dc:coverage)
- Rights Information (dc:rights)
- Subject (dc:subject)

Optional

- Citation (dc:relation)
- Collection Name
- Contributor (dc:contributor)
- Genre (dc:type)
- Keywords or Tags (dc:subject)
- Language of Metadata Record (no dc map)
- Notes (dc:description)
- Publisher (dc:publisher)

Metadata: Crosswalks and interoperability standards – [*this is a little technical – so we shall go slow*]

- Interoperability is the ability of multiple systems, with different hardware and software platforms, data structures, and interfaces, to exchange data with minimal loss of content functionality. There are two approaches to interoperability—cross-system search and metadata harvesting.
- The Z39.50 protocol is commonly used for cross-collection search. The Z39.50 client (called origin) maps search syntaxes to a common set of search attributes for extracting information from Z39.50 server (called target).
- Open Archives Initiative63 is a protocol for metadata harvesting, which allows all partners to translate their native metadata to a common core set of elements and expose those for harvesting.

Metadata Crosswalks

- Metadata crosswalks facilitate the interoperability and exchange of metadata. A crosswalk is a mapping of the elements, semantics and syntax from one metadata schema to those of another.
- It allows metadata created by one community to be used by another group that employs a different metadata standard.
- The Library of Congress' Network Development and MARC Standards Office is developing a framework for working with MARC data in a XML environment.
- This framework is intended to be flexible and extensible to allow users to work with MARC data in ways specific to their needs.
- The framework will contain many components such as schemas, style sheets, and software tools developed and maintained by the Library of Congress.

MARC-XML could potentially be used

- for representing a complete MARC record in XML,
- as an extension schema to METS (Metadata Encoding and Transmission Standard),
- to represent metadata for OAI harvesting,
- for original resource description in XML syntax and
- for metadata in XML that may be packaged with an electronic resource.
- A crosswalk mapping of Dublin Core, MARC 21 and Z 39.50 attributes is illustrated here to make it clear to you.

Crosswalk of DCMES and MARC 21 Bibliographic Data Format

Sl. No	Z39.50 USE Attributes		Dublin Core Elements	MARC21 Fields
	Name	Value		
1	DC-Title	1097	Title	245 $a
2	DC-Contributor	1098	Creator	100, 110, 111, 700, 710, 711 and 720
3	DC-Subject	1099	Subject	600, 610, 611, 630, 650, 653
4	DC-Description	1100	Description	500 –599 excluding 506, 530, 540, 546
5	DC-Publisher	1101	Publisher	260 $a and 260 $b
6	DC-OtherContributor	1106	Contributor	--
7	DC-Date	1102	Date.issued	260 $c
8	DC-ResourceType	1103	Type	655
9	DC-Format	1107	Format	856 $q
10	DC-ResourceIdentifier	1104	Identifier	856 $u
11	DC-SourceIdentifier	1108	Source	786 $o $t
12	DC-Language	1105	Language	008/35-37, 041, 546
13	DC-Relation	1109	Relation	530, 760-787 $o $t
14	DC-Coverage	1110	Coverage	651, 752
15	DC-RightsManagement	1111	Rights	506, 540