DESIDOC Bulletin of Information Technology , Vol. 24, No.4, July 2004, pp. 3-11

A Brief Overview of Metadata Formats

Saji S Nair & V K J Jeevan

Abstract

Metadata is data about data, a term that got wide prominence to denote cataloguing and representation of digital information resources. This paper attempts to present a brief about old and new metadata formats such as Dublin Core, Text Encoding Initiative, Encoded Archival Description, Government (Global) Information Locator Service, Platform for Internet Content Selection, Resource Description Framework, Meta Content Framework, Summary Object Interchange Format, Digital Object Identifier, Serial Item and Contribution Identifier, Uniform Resource Characteristics, and Learning Object Metadata. As electronic information resources are rising and digital library initiatives are getting wide acceptance, knowledge of metadata formats will help our library professionals in adapting their skills in cataloguing, classification, subject heading, key wording, and indexing for better inventory and exhaustive usage of electronic information.

1. INTRODUCTION

Metadata is data about other data and objects, used to describe digitized and non-digitised resources located in a distributed system in a networked environment[1]. Metadata means 'data about data', or machine-understandable information to identify, locate, and/or describe web resources[2]. Equivalent traditional library standards include ISBN and ISSN (identification), shelf mark/call number (location), ISBD and AACR2 (bibliographic description), LC and DDC (subject classification), LCSH (subject headings), and MARC (machine-readable communication format)[2]. Due to the expanding electronic information environment, traditional metadata (library cataloging rules, schemes, and formats) have been expanded to reflect the needs of information discovery and use of information in such a networked environment and thus, rather than 'electronic cataloging rules', the term 'metadata' has been used[1]. This also reflects the major shift in libraries from a largely print and paper dominated information resources to that of electronic and networked information resources. Metadata can include bibliographic information such as that in traditional library catalogs, subject cataloging, such as descriptors, classification designations, abstracts, etc., structural data on the type and size of resources, as well as technical requirements for their use or necessary for access, relationships (thematic, formal, references, citations, etc.), terms and conditions for obtaining and using the resources, etc.[1]

2. METADATA FORMATS

A variety of metadata information schemes have been developed promising greater flexibility in the discovery and access to digitised sources, especially on the web. Several initiatives are underway to explore the possibility of creation and use of metadata primarily concerned for web resources.

This paper purposely avoids MARC, classification and cataloguing schemes, ISBDs, ISBNs, and subject heading lists, etc., from the discussion due to their wide familiarity with library and information professionals in the country.

2.1 Dublin Core (DC)

The Dublin Core (DC) is the most popular and widely accepted standard proposed to describe almost all categories of networked electronic resources. OCLC and the National Centre of Supercomputing Applications (NCSA) developed it jointly and the conceptual framework was developed at a workshop held at Dublin, USA in 1995[3]. The semantics was worked by an international cross disciplinary team comprised of academics, information providers and others interested in the problems of information retrieval.

The DC is a flexible standard having simplicity, extensibility, interoperability, interpretability and international scope, and has a commonly understood semantics. As a pioneer in simple content description, it employs various methods to deploy the metadata such as embedding it within the resource, linking it to a resource or storing it in a database. It provides interoperable metadata standards and specialised metadata vocabularies for describing resources and enable simple, effective and intelligent information retrieval.

The DC has a set of fifteen core descriptive elements relating to content, intellectual property, etc. The applications of DC elements have been designed to cover not only the types of resources in traditional repositories of information, but also on the web. Each of the elements is repeatable and can also have sub-types and sub-object relationships. The element sets describe the simple resource description designed to be used by content creators.

The DC elements are [4]:

(i) Subject and key words: Topic addressed by the work

(ii) Title: Name of the object

(iii) Author or creator: Person(s) primarily responsible for intellectual content of the object

(iv) Publisher: Agent or agency responsible for making the object available

(v) Description: Textual description of content

(vi) Other agent: Person(s), such as editors and transcribers who have made other significant intellectual contribution to the work

(vii) Date: Date of publication

(viii) Object type: Genre of the object, such as novel, poem or dictionary

(ix) Form: Data representation of the object, such as postscript file or windows executable file

(x) Identifier: String or number used to uniquely identify the object

(xi) Relation: Relationship to other objects

(xii) Source: Objects, either print or electronic, from which this object is defined, if applicable

(xiii) Language: Language of the intellectual content

(xiv) Coverage: Spatial locations and temporal duration characteristic of the object

(xv) Rights management: A rights management statement, an identifier that links to a right management statement or an identifier that links to a service providing information about right management for the resource.

The DC standards support cross resource discovery by acting as intermediaries between a large number of community specific formats. The defined list of resource types in DC is strongly oriented to the needs of libraries and other similar agencies. Among library communities, DC has generated special interest because of the extensibility and easy element transfer to machine readable catalogues. Several cooperative/ collaborative digitisation projects have been taking advantage of the elements of DC with some modifications to create their base records.

2.2 Text Encoding Initiative (TEI)

It is an international research effort established in 1987, intended to produce a community-based standard for encoding and interchange of texts. The original phase of the effort was sponsored by three of the most important scholarly associations such as the Association for Computers and Humanities (ACH), the Association of Computational Linguistics (ACL) and the Association of Literary and Linguistic Computing (ALLC) and included scholars all over the world from nearly all disciplines of the humanities. Major support for the project has come from the US National Endowment for the Humanities (NEH), Directorate XIII of the Commission of the European Communities (CEC/DG-XIII), the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada.

In December 2000, a new non-profit corporation called the TEI Consortium was set up to maintain and develop the TEI standard. The Consortium has executive offices in Bergen, Norway, and hosts at four universities: University of Oxford Research Technologies Service, Brown University Scholarly Technology Group, University of Bergen Humanities Information Technologies Research Programme, and University of Virginia Electronic Text Center and the Institute for Advanced Technology in the Humanities[5]. The TEI Consortium identifies three categories of participation: voting membership, meant for companies, institutions or projects; non-voting subscription, which is open to personal individuals only, and sponsorship, for individual or corporate sponsors.

There are currently over 100 projects using TEI[5]. As TEI is expressed in SGML or XML, any software supporting those standards supports TEI. However, a list of TEI specific software tools is available on the TEI web site[5].

The TEI guidelines and the SGML-based TEI DTD (Text Encoding Initiative Document Type Definition) are the result of the twentieth century's most important international standardization effort for humanities related data and beyond.

The TEI guidelines specify that every TEI text must be preceded by a TEI header that describes the text[1]. The TEI headers are used as a means of bibliographic control and the header segment contain rich tag sets, which can sufficiently support library cataloguing practice with AACR II rules and authority control. The initiative provides greater flexibility in creating TEI headers for author, publisher of electronic text and archive administrators. To get the best out of the TEI, we need to customize it to suit our requirements, which requires some knowledge of the whole of the TEI, if only in order to know what to throw out. TEI Lite is a specific customization designed for the core TEI constituency, which has proved very popular.

Some elements in the TEILITE.DTD are listed below.

ABBR	'Abbreviation or acronym'
ANCHOR	'Identifiable point in a text'
AUTHOR	'Author or authors names'
AUTHORITY	'Release authority name'
AVAILABILITY	'Text availability'
BACK	'Back matter'
BIBL	'Loosely-structured bibliographic citation'
BIBLSCOPE	'Scope of bibliographic citation'
BODY	'Body of a text, excluding front or back matter'
BYLINE	'Primary statement of responsibility'
CATDESC	'Description of single category'
CHANGE	'Change or revision'
CLASSCODE	'Classification code'
CREATION	'Creation information'
DATE	'Date'
DATELINE	'Place, date, time, etc., used as heading'
DISTRIBUTOR	'Agency responsible for distribution'

DOCAUTHOR	‘Author’s name as given on title page‘
DOCDATE	‘Date as given on titlepage‘
DOCEDITION	‘Edition statement as given on title page‘
DOCIMPRINT	‘Imprint statement as given on title page‘
DOCTITLE	‘Title as given on title page‘
EDITION	‘Edition description‘
EDITIONSTMT	‘Edition information‘
EDITOR	‘Secondary statement of responsibility‘
EDITORIALDECL	‘Details of editorial practice‘
EXTENT	‘Approx size of text as stored on some medium‘
FIGDESC	‘Description of inserted graphic or figure‘
FIGURE	‘Inserted graphic or figure‘
FILEDESC	‘Full bibliographic description of electronic text‘
FUNDER	‘Name of agency responsible for funding project or text‘
IDNO	‘Standard number used to identify bibliographic item‘
IMPRINT	‘Information concerning publication of text‘
KEYWORDS	‘Keywords describing topic or nature of a text‘
LANGUAGE	‘Language used in a text‘
NAME	‘Name of person, place, organisation, etc.‘
NOTE	‘Annotation‘
PRINCIPAL	‘Name of principal researcher‘
PUBLISHER	‘Name of publisher or distributor‘
PUBPLACE	‘Place of publication‘
RESPSTM	‘Statement of responsibility‘
REVISIONDESC	‘Revision history‘
SPONSOR	‘Name of sponsoring organisation‘
TEIHEADER	‘The header of a TEI-conformant document‘
TERM	‘Technical term‘
TIME	‘Phrase defining time of day‘
TITLE	‘Title of a work‘
TITLEPAGE	‘Title page within front or back matter‘
XPTR	‘Pointer to an external document‘
XREF	‘Reference to an external document‘

2.3 Encoded Archival Description (EAD)

The EAD Document Type Definition (DTD) is a standard maintained by the Network Development and MARC Standards Office of the Library of Congress (LC) in partnership with the Society of American Archivists, for encoding archival finding aids using the SGML[6]. The archival arrangement and description establish physical and intellectual control over archives and manuscripts enabling users to find the records they need. EAD is the SGML/XML based DTD that archives, libraries and museums are using to create, store and distribute descriptions of their collections[7]. Basically EAD is a richest level data communication format and typically applied to archival materials.

Some EAD elements are listed below:

<Abstract> -	Abstract
<accessrestrict> -	Conditions governing access
<address> -	Address
<archdesc> -	Archival description
<archref> -	Archival reference
<author> -	Author
<bibliography> -	Bibliography
<c> -	Component (Unnumbered)
<c01> to <c12> -	Component (first level to twelfth level)
<corpname> -	Corporate name
<creation> -	Creation
<dimensions> -	Dimensions
<ead> -	Encoded Archival Description
<eadgrp> -	EAD group
<eadheader> -	EAD header
<eadid> -	EAD identifier

<edition> -	Edition
<editionstmt> -	Edition statement
<extent> -	Extent
<geogname> -	Geographic name
<head> -	Heading
<head01> -	First heading
<imprint> -	Imprint
<index> -	Index
<langmaterial> -	Language of the material
<language> -	Language
< materialspec> -	Material specific details
<note> -	Note
<notestmt> -	Note statement
<persname> -	Personal name
<physdesc> -	Physical description
<phystech> -	Physical characteristics and technical requirements
<publicationstmt> -	Publication Statement
<publisher> -	Publisher
<relatedmaterial> -	Related material
<resource> -	Resource
<revisiondesc> -	Revision description
<seriesstmt> -	Series statement
<sponsor> -	Sponsor
<subject> -	Subject
<subtitle> -	Subtitle
<title> -	Title
<titlepage> -	Title page
<titlestmt> -	Title statement
<unitdate> -	Date of the unit
<unitid> -	ID of the unit
<unittitle> -	Title of the unit
<userestrict> -	Conditions Governing use

The description was developed by archivists in USA to support the inventories and registries on the web and its flexibility has helped it to spread throughout the world. EAD was originally begun with a project initiated by the University of California, Berkeley library in 1993. Now it forms a standard for machine-readable finding aids such as inventories, registries and other documents created by archives, libraries, museums and manuscript repositories.

However this is not regarded as a system for collection management activities such as transfer of ownership, conservation, exhibition, storage or technical processing of materials. In EAD, the collections are described in an intellectual sense rather than a physical sense. Even though EAD fully supports archival description practices, it lacks robust mechanism for web resource discovery that DC metadata standard provides.

2.4 Government (Global) Information Locator Service (GILS)

It is a tool for the identification of US Government information resources and is now being adopted in other countries. GILS provides users with a means of finding information located in local and remote systems and adopts the concept of topic trees to classify the varied repository of information. It has a decentralised collection of systems containing database of GILS records describing location and access information for publicly accessible information resources [8].

Databases, clearinghouses and catalogues are the type of resources described. GILS record for these resources are not resources themselves, but are descriptions of the resources which inform users what information is available, where it is located and how it can be accessed. In cases in which the resource exists in an electronic form, a direct link to that resource from the GILS record will usually be available[9]. In short, GILS has similarity to the ‘bibliography of bibliographies’ concept and with many accessible links, it may also function as a virtual library.

Some of the ‘Repeatable’ GILS core elements are Contributor, Language of resource, Time period, Availability, Resource description, Cross reference, etc., and ‘Not Repeatable’ core elements are Title, Date of publication, Place of publication, Abstract, Spatial domain, Point of contact, Schedule number, Control identifier, Original control identifier, Record source, Language of record, Date of last modification, and Record review date.

2.5 Platform for Internet Content Selection (PICS)

Platform for Internet Content Selection (PICS) comprise of a set of metadata specifications proposed by World Wide Web Consortium (W3C) Working Group which enable people to distribute metadata about the content of digital material in the form of labels. They present information in simple, computer readable form. A label consists of a service identifier (URL chosen by the rating service as its unique identifier), label options (additional properties of the document being rated as well as properties of the rating itself, such as the time the rating was conducted), and a rating (set of attribute-value pairs that describe a document along one or more dimensions).

The aim of this platform is to provide Internet users the ability to select resources effectively based on metadata description, which will be carried with resources. When an end-user asks to see a particular URL, the software filter fetches the document but also makes an inquiry to the label bureau to ask for labels that describe that URL and depending on what the labels say, the filter may block access to that URL.[10] PICS devised a set of standards that facilitate self-rating (content providers voluntarily labeling the content they create and distribute), third-party rating (independent labeling services associating additional labels with content created and distributed by others) and ease-of-use (ratings and labels from a diversity of sources helps the users to control the information they receive).

2.6 Resource Description Framework (RDF)

The Resource Description Framework (RDF) integrates a variety of applications from library catalogs and worldwide directories to syndication and aggregation of news, software, and content to personal collections of music, photos, and events using XML as interchange syntax and the RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the web.[11]

A set of conventions was developed by the W3C, a standard body for web technology working for standardization of metadata formats. The goal is to support interoperability of metadata describing any item that can have Uniform Resource Identifier (URI). Here a resource is described through a collection of properties (each with a Property Type and Value) called an RDF description and any resource can be described with RDF as long as the resource is identifiable with a URI. RDF has proposed a robust and flexible architecture for supporting metadata and the XML is the encoding scheme.

RDF uses the key concepts of graph data model [subject, object, and predicate (or property that denotes the relationship between subject and object) triple] with subject and object nodes and graph pointing towards object, URI-based vocabulary (A node may be a URI with optional fragment identifier, a literal or blank with no separate form of identification and a URI reference may be used as a predicate), datatypes (to represent values such as integers, floating point numbers and dates), literals (to identify values such as numbers and dates by means of a lexical representation), XML serialization syntax, expression of simple facts [a more complex fact is expressed using a conjunction (logical-AND) of simple binary relationships], and entailment (an RDF expression A is said to entail another RDF expression B if every possible arrangement of things in the world that makes A true also makes B true).[12]

Resource description, site maps, content rating, electronic commerce, privacy protections, etc., are the core areas. RDF provides a common framework for representing metadata about Web resources such as title, author, modification date of web pages, copyright and licensing information about a web document or the availability for some shared network.[11] Since PICS is a predecessor to RDF, it is an explicit requirement of RDF that it must be able to express anything that can be expressed in a PICS label.

2.7 Meta Content Framework (MCF)

The Meta Content Framework (MCF) provides a system for representing a wide range of information about content. The content targeted includes web page, gopher and ftp files, e-mail and structured (i.e., relational and object oriented) databases, etc. MCF is not intended to be an extension of Markup languages such as HTML that can be used to hold embedded metadata. Instead it provides a format for holding the metadata externally to the content described.[13]

The MCF is an open format for representing information about the content. It provides information about information by attaching properties to objects. For example, a web page could have one property that give its size, another that gives its URL and another that identifies the person who maintains it, etc.[14]

2.8 Summary Object Interchange Format (SOIF)

It is the internal record format of the Harvest, the information discovery access system addressing the problems relating to information access via web. SOIF is based on simple attribute-value pair elements.

A single SOIF stream can contain multiple SOIF 'templates', each with URL for the resource it refers to and a number of different elements for holding other metadata. Each element has an attribute name, the length of the value in brackets, a colon delimiter and then the value itself.[15] It is the structured indexing format for more complex search applications like image and audio searching.

2.9 Digital Object Identifier (DOI)

It is an identification system for intellectual property in the digital environment. The A merican Association of Publishers (AAP) and the Corporation for National Research Initiative (CNRI) jointly developed it in 1996.

The goals are to provide a framework for managing intellectual content, link customers with publishers, facilitates electronic commerce and enable automated copyright management.[16] DOI is presently controlled by a not-for-profit organisation called International DOI Foundation (IDF).

The DOI system is primarily meant for the effective rights management and digital commerce. Linking other systems like ISBN, ISSN, URL, etc., are the areas of concern and the key issue is metadata definition. The DOI utilises the DC metadata for describing the object itself as well as additional publisher oriented information such as price and purchase conditions down to the level of individual article, diagram or graphic. As a unique, persistent, managed international public identifier, the system helps libraries to identify, locate, and link the items. Users need not worry about expired URLs because the DOIs will always point to the current location of digital objects.

The DOI system comprises of three parts such as an identifier, a directory system and a logical database. The identifier has a prefix code supplied by the DOI registration agency and a suffix code usually assigned by the concerned publishing industry. The system uses a central directory to rectify problems relating to change of ownership and location of digital content. The information about the object identified is stored in the database of DOI system.

2.10 Serial Item and Contribution Identifier (SICI) (ANSI/NISO Z39.56)

Serial Item and Contribution Identifier (SICI) was defined by the US based Serials Industry Advisory Committee (SIAC) in 1991. It is intended primarily for use by those members of the bibliographic community involved in the use or management of serial titles and their contributions.

It can identify a serial issue or item, a particular contribution by reference to its physical appearance in an issue item and a contribution without a physical appearance. It identifies serial issues and articles uniquely regardless of distribution medium.

2.11 Uniform Resource Characteristics (URC)

The Universal Resource Locator (URL) is the networked extension of a standard file name concept. As one of the widely developed identification scheme for resources on the web, URL identifies the location, place of a resource identified by a Universal Resource Name (a logical name of resources in the Internet). The URL describes the current location of resource and is subject to change if the resource is moved. So it is not regarded as a persistent identification for electronic document. The electronic address of URL specifies a communication protocol with host domain/server, directory path, file name and file type in a specific sequence. Precisely, the URIs that refers to objects with existing protocols is known as URLs.

To minimise the location dependence problems and a possible solution to the unpredictable mobility of internet resources, Persistent Uniform Resource Locator (PURL) is proposed. Functionally PURL is a URL, but instead of pointing directly to the location of an internet resource, it points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns the URL to the client.

The client can then complete the URL transaction in the normal fashion. A PURL can be associated with any given resource. PURLs can be embedded in catalogue record, finding aids and other types of metadata used for linking to objects.

Internet Engineering Task Force (IETF), the standard setting body for internet development, recognised the limitations of URL schemes and initiated a working group to establish a parallel scheme called Universal Resource Locator (URL) is to provide a unique persistent identification for a resource that is not dependant on location. The standard requires URN registries to avoid duplication and resolution system to map to the location of the resource. URN is not a standard as yet and is still under development.

Resources are named by URN and are retrieved by means of URL. The role of URC is to make the binding between the URN of a resource and its URLs. In addition URC can contain metadata about the resource for the purpose of discovery, conveying usage restrictions, etc. URC will also contain metadata about a particular URN. The purpose or function of URC is to provide a vehicle or structure for the representation of URIs and their other associated meta-information.

The URC is developed in conjunction with URNs as a means of describing internet accessible resources. URCs have a set of values, which may include authorship, publisher, data type, and copyright statement, etc. Not implemented so far.

The URI encompasses a group of technologies developed by the IETF, for naming, addressing and describing web resources. It consists of short strings that identify resources in the web such as document, images, downloadable files and other resources. As an extendable form of approach, URI currently includes URL, URN and URC.

2.12 IEEE-Learning Object Metadata (LOM)

IEEE Standard for Information Technology—Education and Training Systems—Learning Objects and Metadata specifies the syntax and semantics of LOM, the attributes required to fully/adequately describe a learning object. Learning object is any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning. The LOM standards will focus on the minimal set of attributes needed to allow learning objects to be managed, located, and evaluated. The standards will accommodate the ability for locally extending the basic fields and entity types, and the fields can have a status of mandatory or optional. Relevant attributes to be described include type of object, author, owner, terms of distribution, and format. It may also include pedagogical attributes such as; teaching or interaction style, grade level, mastery level, and prerequisites. It is possible for any given learning object to have more than one set of LOM. The standard will support security, privacy, commerce, and evaluation, but only to the extent that metadata fields will be provided for specifying descriptive tokens related to these areas; the standard will not concern itself with how these features are implemented.[17]

3. CONCLUSION

Awareness about metadata formats will help the library and information professionals to take effective measures about properly identifying, representing and disseminating digital information. Since there are many standards to choose from, there is little progress in implementing any one of them even in countries with sufficient electronic resources and associated finding information. However, the acceptance of DC as an important metdata format is changing the scenario. As electronic information resources are rising and digital library initiatives are getting wide acceptance, it is the right time for librarians in the country to be involved in activities concerning how to mature their traditional metadata skills of cataloguing, classification, subject heading, key wording, and indexing to represent electronic information for their proper inventory and better usage.

REFERENCES

1. Rusch-Feja, Diann. Metadata: Standards for retrieving www doc u ments (and other digitized and non-digitized sources). *Library and Information Services in Astonomy III*, 1998, **153**. http://www.stsci.edu/stsci/meetings/lisa3/ruschfejad.html
2. Susan Haigh. A glossary of digital library standards, protocols and formats. Network Notes # 54.
3. Dublin Core Metadata Initiative (DCMI). http://www.dublincore.org
4. El Sherbini, Magda. Metadata and the future of cataloging. *Library Review,* 2001, **50** (1), 18-27.
5. Text Encoding Initiative (TEI). http://www.tei-c.org
6. http://www.loc.gov/ead/
7. Keisling, Kristi. Metadata, metadata everywhere but where is the hook? *OCLC Systems and Services*, 2001, **17**(2), 84-89.
8. Government Information Locator Service (GILS). http://www.gils.net
9. What is government information locator service? www.epa.gov/earth100/docs/whatis.html
10. http://www.w3.org/PICS/
11. Resource Description Framework (RDF). http://www.w3.org/RDF/
12. Resource Description Framework (RDF): Concepts and abstract syntax. W3C Working Draft 05 September 2003. http://www.w3.org/TR/rdf-concepts/
13. Meta Content Framework (MCF). http://www.ifla.org/II/metadata.htm#mcf
14. Bray, Tim & Guha, R.V. An MCF tutorial. http://www.w3.org/TR/NOTE-MCF-XML/MCF-tutorial.html
15. The UKOLIN Metadata group. Summary Object Interchange Format (SOIF). www.ukoln.ac.uk/metadata/desire/overview/rev_20.htm
16. International DOI Foundation (IDF). http://www.doi.org
17. IEEE Learning Technology Standards Committee (LTSC), IEEE P1484.12 Learning Object Metadata Working Group. http://ltsc.ieee.org/wg12/s_p.html

Contributors: **Sh Saji S Nair**, Library Assistant, Regional Cancer Centre, Medical College Campus, Trivandrum, Kerala - 695 011, e-mail: saji@rcctvm.org

Sh V K J Jeevan, Assistant Librarian, Indian Institute of Technology, Kharagpur, West Bengal - 721 302, e-mail: vkjj@rediffmail.com