Metadata Standards for Textual and Multimedia Content

ARD Prasad Devika P Madalli

Abstract

Metadata, data about data, describes objects of various multimedia content and types. One of the main functions of metadata is aiding in retrieval of the objects that it describes. However, with the emergence of semantic web technologies the role of metadata is changing. Paper discusses some of the important metadata schema. It highlights issues of 'glut' by way of a plethora of metadata that are emerging. In the given scenario, it is essential to devise a method for achieving interoperability among similar metadata standards.

Introduction

Metadata can be used as a means of describing any object. Hence there are plethora of metadata schema for various kinds of objects. Broadly, metadata schema can be classified as Bibliographic metadata schema and non-bibliographic metadata schema. This kind of classification is based on the assumption that the origins of metadata can be traced back to library cataloging. Indeed Dublin Core aimed at description of documents. The bibliographic metadata schema can be further sub-divided into Textual and Multimedia media related metadata schema. The question arises whether multimedia related metadata can be considered as bibliographic metadata. Libraries always hosted resources of various media and the connotation used was 'multimedia library' to include materials such as maps, atlases, software which can appear in various media like tapes, CDs, DVDs etc. In fact, MARC21 facilitates description of various multimedia objects.

The semantic web technology has evolved other schema that include any kind of resource. The Resource Description Framework (RDF) is developed and is extended to Web Ontology Language (OWL) and Simple Knowledge organization System (SKOS). Again, it should also be kept in mind Dublin Core can be expressed in RDF. Many of the metadata schema can be expressed in RDF or OWL formats. Technically, there can be metadata for documents, people, institutions, software, events, processes etc. This typically reflected in various active group like:

TV-Anytime Forum aims to provide specifications for audio–visual and other services for digital storage (http://www.tv-anytime.org/). NewsML aims to provide structured framework for multimedia news. (http://www.newsml.org/)

This paper confines itself to Bibliographic metadata schema, though OWL and SKOS are being used to represent thesauri and classification systems, which play an important role in web based information retrieval and information services. Again, many of these schema can be expressed in terms of RDF and OWL to facilitate semantic web.

The most exhaustive and complex metadata schema is MARCXML and the most simple metadata schema is Dublin Core. The both represent two extreme positions and in between these extremes one may find Qualified Dublin Core (QDC), MODS, ETDMS, FRBR, ONIX etc. The following sections present a brief information about some of the metadata schema.

The Dublin Core Metadata (DCMI) *(www.dublincore.org)* deemed as the generic metadata standard, Dublin Core Metadata Schema has 22 (15 main) elements. DC elements are often used in conjunction with other refinements where more detailed or specific descriptions are required. The basic elements are refined using qualifiers and this is known as **Qualified Dublin Core** (QDC). In QDC qualifiers are used to refine the semantics of the elements in ways that may be useful in resource discovery.

MODS : Metadata object description schema (MODS) *(http://www.loc.gov/standards/mods/)* is maintained by the Library of Congress. MODS is a schema for a bibliographic element set that may be used for library applications among other applications. It is a descriptive metadata schema that is a derivative of MARC 21 and intended to either carry selected data from MARC 21 or enable the creation of original resource description records. It includes a subset of MARC fields and uses language based tags rather than the numeric ones used in MARC 21 records.

FRBR : Functional Requirements for Bibliographic Records (FRBR) was developed by the IFLA study group. FRBR deals with structure and relationships of bibliographic and authority records, and also provides a more precise vocabulary to help catalogers and

system designers in meeting user needs. It is designed as an entity-relationship model that provides a generalized view of the bibliographic universe, intended to be independent of any cataloging code or implementation.

ONIX : Online Information eXchange (ONIX) was originally devised to simplify the provision of product information to online retailers of books. ONIX is used by book publishing industry for providing information to retailers and wholesalers about their products. ONIX is also used for storing and sharing book information. ONIX was originally developed by the Association of American Publishers and EDItEUR. ONIX elements give book information such as price, availability & supplier, blurb, reviews & extracts and detail on formats

TEI : The Text Encoding Initiative (TEI) is an international cooperative research effort, the goal of which is to define a set of generic Guidelines for the representation of textual materials in electronic form. The Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange (referred to as the *TEI Guidelines*) were first published in 1994. TEI scheme enhances the content by facilitating the integration of multimedia technology into electronic resources. This is done by providing a description of information that is independent of realization or media of the resource. TEI explicitly represents the different features of text, which in turn makes it easy to manipulate the processing through computer programs. A set of tags or markers are specified, which may be inserted in the electronic representation of the text, in order to mark the text structure and other textual features of interest.

Domain Specific Metadata:

AGRISAP : The AGRIS Application Profile (AGRIS AP) *(www.fao.org/ docrep/008/ae909e/ ae909e02.htm)* is a metadata standard created specifically to enhance the description, exchange and subsequent retrieval of agricultural Document-Like Information Objects (DLIOs). It allows sharing of information across dispersed bibliographic systems and provides guidelines on recommended best practices for cataloguing and subject indexing in agriculture domain.

Multimedia Metadata:

VRA Core : VRACore (http://www.vraweb.org/projects/vracore4/) was developed by the Visual Resources Association's Data Standards Committee as a data standard for the cultural heritage resources. It is defined as, "A single element set that can be applied as many times as necessary to create records to describe works of visual culture as well as the images that document them" (http://www.vraweb.org/projects/vracore4/) . The element set provides a categorical organization for the description of works of visual culture as well as the images that document them. The Core elements pertain to only two types of records - the work (resource), the image and other types of records such as authority records may be included in a database structure. VRA Core is structured so it can be used in a variety of databases and encoding schemes.

MPEG Metadata

The MPEG-7 & MPEG-21 community aims at collaboration through multimedia metadata interoperability *(http://www.multimedia-metadata.info/)*. The MPEG-7 **c**alled the *Multimedia Content Description Interface*, provides a tool set for completely describing multimedia content. MPEG-7 is designed to be generic and not targeted to a specific application. MPEG-21 describes a standard that defines the description of content and also processes for accessing, searching, storing and protecting the copyrights of content.

MPEGs: MPEG is an open standard for encoding the movies, where as the MPEG-2 extended its coverage to Video-CD, DVD, digital television, digital audio broadcasting (DAB) and MP3 (MPEG-1 Audio layer 3) players and recorders. The MPEG-4 aims at more structured content to enable interactivity. The MPEG-4 standard allows users to play with, re-use, and access audiovisual content. It is MPEG-7 which attempts to describe the content (metadata) of audio-visual data, as metadata provides a powerful solution for quickly and efficiently identifying, searching, filtering of audiovisual information. However to address interoperability issues and to provide a broader framework, the MPEG-21 envisioned to define an open multimedia framework to enable the transparent and augmented delivery and consumption of multimedia resources across a wide range of networks and devices used by different communities.

DIDL (Digital Item Declaration Language): "The basic architectural concept in MPEG-21 is the Digital Item. Digital Items are structured digital objects, including a standard representation, identification and metadata. They are the basic unit of transaction in the MPEG-21 framework. More concretely, a Digital Item is a combination of resources (such as videos, audio tracks, images, etc.), metadata (such as descriptors, identifiers, etc.), and structure (describing the relationships between resources). (http://xml.coverpages.org/mpeg21-didl.html).

DSpace digital repository supports DIDL though OAI-PMH protocol, so that metadata of any multimedia items of DSpace based repository can be harvested by OAI service providers.

METS : Metadata Encoding and Transmission Standard (METS) does not include descriptive metadata. However, it can accommodate any descriptive metadata schema like Dublin Core, VRA Core etc. In brief, METS is a standard for transmitting and or exchanging digital objects. It forms a standard basis for providing end users with the ability to view and navigate digital content and its associated metadata.

METS is intended to provide a standardized XML format for transmission of complex digital library objects between systems. As such, it can be seen as filling a role similar to that defined for the Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP) in the Reference Model for an Open Archival Information System.

Features of METS

METS allows

- to point to external descriptive metadata or for including metadata internally. It provides a way for linking this metadata to the digital content of the digital object.
- to link to external administrative metadata or for including metadata internally. It provides a way for linking this metadata to the digital content of the digital object.
- to specify the structure of a digital object i.e. how the files and parts of files of a digital object are linked.

- to provide a way for linking digital content with external software capable of disseminating that content, as well as an interface file that defines the specific dissemination and the required parameters for each through its behaviors section.

Some Observations:

As Dublin Core (DC) is an approved W3C standard, various harvesting services are compliant with DC. Though, more granular schema are available, as they are yet to prove their distinct advantages, it may take more time for these to mature, so that they can be incorporated into harvesters and other web-based information systems and services. For example, DSpace allows exposing metadata through OAI-PMH in various schema like DC, QDC, DIDL, METS, MODS. But we do not have many harvesting software to accept metadata schema other than DC.

The purpose of metadata is similar to that of cataloging, that is, to provide access to information. This purpose can be extended to harvesting (shared cataloging), filtering information and to a host of information services. In order to achieve various objectives, it is of foremost importance to provide interoperability. Though metadata crosswalk (retro-conversion) can be achieved using XSLT, any conversion results in loss of data and granularity.

The next probable issue is that whether different metadata schema can augment each other to enrich metadata that are described in different schema representation about a particular digital object, using the unique URI representing one object and described in various schema. Digital Object Identifiers (DOI) and CNRI handle are presently used, they are mostly for digital documents like journal articles and full text digital objects in various digital repositories.

The next big question is how to evolve a unique URI, so that whatever be the description, the machines should be able to gather metadata about a particular URI and merge the various metadata descriptions to achieve a comprehensive description about an object. The above mentioned are research issues that when solved will help achieve true interoperability among various metadata standards without having to impose one on another.

Conclusion

The library science field from its origins realized the importance of cataloging and classification in information retrieval. The philosophy behind cataloging was extended in the context of computerization to various bibliographic databases like online databases and CD based databases. As traditional classification was more concerned with coding to facilitate arrangements of documents on a shelf, it appeared to have no relevance in the context of automation. However, if we visualize classification sans notation and use classificatory techniques in subject indexing, the importance of classification is more obvious in relating objects with each other. Thus, classification reappeared in semantic web technology in the form of ontologies. If we consider a piece of information as a object, cataloging (metadata) can be effectively used to describe an information object and classification (ontology) can be used to find relation of an information object with other information object. Ultimately, all these metadata schema are to be expressed in RDF or its extensions like OWL.

In the era of libraries, cataloging and classification proved their ability in bibliographic management. However, the Web and digital libraries that are emerging on the Web pose an altogether novel challenges, as the Web hosts more complex forms of information in different formats. Though the origins of metadata can be traced back to library cataloging, it is being extended to not only of documents, but also to persons, institutions, services, events, products etc. Similarly, library classification has evolved into web ontology. Web ontology is not mere representation of hierarchical relations of subject domains as in the case of thesauri and library classification schemes. It is expressing the relations in a kind of predicate logic where inference engines can be used deduce more specific information.

It is advantageous to have metadata of information objects, however, having too many metadata schema pose unwieldy issues. This is similar situation to that of a plethora of MARCs adopted by various countries across the world in 80s and 90s. It is good to have metadata to make information available on the web more meaningful, however, having too many metadata schema will lead to major interoperability issues. Unless, the heat and dust of many schema settles, it is difficult to realize semantic web.

The evolution of an unified approach to metadata and ontology is of paramount importance. In the evolutionary process, some of the existing schema may further developed and refined and some may get extinct.

References

1. Dublin Core. http://dublincore.org
2. FRBR http://www.loc.gov/cds/downloads/FRBR.PDF
3. Metadata Transimission and Encoding Standard (METS): http://www.loc.gov/standards/mets/
4. Metadata Object Description Schema (MODS): http://www.loc.gov/standards/mods/
5. MARCXML: http://www.loc.gov/marc/marcxml.html
6. Introduction to RDF: www.w3.org/TR/NOTE-rdf-simple-intro-971113.html
7. Web Ontology Language (OWL): http://www.w3.org/RDF/
8. ONIX ONline Information eXchange: Metadata Reference Guide: http://libraries.mit.edu/guides/subjects/metadata/standards/onix.html
9. TEI (Text Encoding Initiative): Metadata Reference Guide: http://libraries.mit.edu/guides/subjects/metadata/standards/tei.html
   VRA Core: http://www.vraweb.org/projects/vracore4/
10. MPEG-7 Overview: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
11. EAD – Encoded Archival Description: Metadata Reference Guide: http://libraries.mit.edu/guides/subjects/metadata/standards/ead.html

About Authors

Dr. ARD Prasad, Associate Professor, DRTC, ISI, Bangalore.

Dr. Devika P. Madalli, Lecturer, DRTC, ISI, Bangalore.