DESIDOC Journal of Library & Information Technology, Vol. 35, No. 3, May 2015, pp. 193-199 DOI: 10.14429/djlit.35.3.8074


# Metadata Diversity, Interoperability and Resource Discovery Issues and Challenges

P. Ramesh*, J. Vivekavardhan** and K. Bharathi***

*Department of Library and Information Science*
*Osmania University, Hyderabad*
*E-mail: *ramesh.parichi@gmail.com, **jvk_dn@yahoo.com, ***bharathikandimalla@yahoo.com*

**ABSTRACT**

Information resources are available in various kinds of media and forms. To describe them there exists number of diverse metadata standards and schema. Metadata is crucial for preservation and archiving, organisation, resource discovery and information retrieval across platforms. As one metadata standard cannot be applicable for all the emerging media and document formats, a combination of them is used. In this context the present paper presents an overview on types of metadata standards and schemas, and also discusses on the issues and challenges in metadata creation, management, interoperability, and resource discovery.

**Keywords:** Metadata, metadata diversity, diverse digital forms, interoperability, resource discovery, crosswalking, registries

## 1. INTRODUCTION

The World Wide Web is repertoire of information resources available in diverse formats and is growing at an exponential rate. This huge availability of information resources led to the problem of finding potentially useful documents. Metadata, the structured data about information sources, is a key to retrieve relevant information that provides meaningful access points for the searchers. Paradoxically, the metadata is both a key and an obstacle in finding relevant information. In the recent years the content in the scholarly materials has expanded beyond the text realm to diverse resources such as multimedia, artistic and creative works, grey literature and research data including trends and reviews. The scope and purpose of metadata differ among different knowledge communities and institutions such as information managers and scientists, IT professionals, government agencies, learning institutions, cultural and heritage organisations, publishers, businesses, legal community and many more. There exists different metadata schemas and standards that are best suited to their purposes and priorities to meet their needs. The present paper discusses the diverse metadata schemas and standards, interoperability and resource discovery, issues and challenges.

## 2. METADATA

Metadata is the descriptive and classification information about digital object or digital resource such as web page, a computer file, an image, multimedia document, etc. It describes the who, what, when, where, why, and how about a data set or resource. Metadata gives complete description about a digital object, and is crucial for preserving and sharing such resources. Besides, providing information it helps the search engines and other discovery channels in discovering and reporting the usefulness of the resources thereby helps the searcher in selecting the best. While identifying the core metadata elements, the focus should also be made on how to share this data with other collections, catalogues and systems to maximise the potential use and re-use of resources.

According to Natasa a metadata framework[1] can be viewed as having five key components: (i) Schema-deals with elements and categories of information to be chosen for recording, (ii) Vocabulary-collection of the specific words and their values chosen for the defined categories, (iii) Conceptual model-framework which describes how different concepts of information are related to each other (iv) Content standard-practical way of recording the specific information (e.g. vocabularies) are entered within metadata schema categories (e.g. cataloguing cultural objects), and (v) Encoding-concerned with the way the metadata is presented (e.g. XML).

### 2.1 Metadata Standards and Schemas

To describe particular information resources, the user communities develop metadata schemas[2],

Received 14 October 2014, revised 3 March 2015, online published 22 April 2015

while practicing they obtain consensus following all the formalities such as submissions, approvals and publishing of new elements, and when ratified by professional organisations they become metadata standards.

According to NISO, a metadata schema[3] consists of a set of elements designed for a specific purpose, such as describing a particular type of information resource. Report of the American Library Association Committee on Cataloguing : Description and Access, Task Force on Metadata (CC:DA 2000) states that a metadata schema[4] provides a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of the discipline through the creation of an information system that will assist the identification, discovery, and use of information within that discipline.

The word 'schema' usually regarded as an 'element set' refers to an entire entity including the semantic and content components, as well as the encoding of the elements with a markup language such as SGML (Standard Generalized Markup Language) and XML (Extensible Markup Language). A metadata element set has two basic components[5]: (i) Semantics-definitions of the meanings of the elements and their refinements (ii) Content-declarations or instructions of what and how values should be assigned to the elements.

### 2.2 Need for Diverse Metadata Standards and Schema

Recently, it has been observed that many scholars like to publish their research work in online e-journals, web forums and scientific datasets. The communities and content providers in the areas of sciences, social sciences, e-commerce and business management, medical sciences, life sciences, biodiversity, space sciences, museums, etc., are on constant lookout for refinements and agreements in the semantics of specialised metadata to enable new formats internationally. As libraries and information centres, museums digitise cultural heritage information, must create metadata to organise and manage it in the consumer world, metadata exists in the form of ratings and reviews; metadata to support medical sciences and rights management will be an essential part. A metadata type which is applicable for all kinds of information resources, knowledge communities, and applications does not exist.

### 2.3 Types of Metadata Standards and Schemas

With so many diverse formats of documents or digital objects, type of user community, type of institutions existing, unfortunately, a unique definitive metadata standard does not exist. Metadata schemes are the sets of metadata elements designed for a specific purpose. Each scheme has a limited number of elements, with meaning of each element specifying the content/syntax rules. The following is not the exhaustive list of the metadata standards and schemas:

(1) Categories for the Description of Works of Art (CDWA)

(2) Describing Archives: A Content Standard (DACS)

(3) EXIF (Exchangeable Image File Format)

(4) e-GMS (e-Government Metadata Standard)

(5) ISO 19115: 2003(E)-Geographic Information: Metadata

(6) PREMIS (PREservation Metadata Implementation Strategies)

(7) Resource Description and Access (RDA)

(8) Machine Readable Catalogue (MARC)

(9) Dublin Core Metadata Initiative (DCMI)

(10) Text Encoding Initiative (TEI)

(11) Metadata Encoding and Transmission Standard (METS)

(12) Metadata Object Description Schema (MODS)

(13) MIDAS Heritage: the UK Historic Environment Data Standard

(14) Encoded Archival Description (EAD)

(15) IEEE Learning Object Metadata (LOM)

(16) Functional Requirements for Bibliographic Records (FRBR)

(17) NISO Technical Metadata for Digital Still Images

(18) PBCore (or the Public Broadcasting Metadata Dictionary)

(19) SMPTE (Society of Motion Picture Technical Experts) Data Dictionary

(20) SPECTRUM

(21) TV-Anytime

(22) UK LOM Core

(23) XMP (Extensible Metadata Platform)

The following are the metadata encoding schemas:

(1) HTML (Hyper-Text Markup Language)

(2) SGML (Standard Generalized Mark-up Language)

(3) XML (Extensible Mark-up Language)

(4) RDF (Resource Description Framework)

(5) MIME (Multipurpose Internet Mail Extensions)

(6) IPTC (International Press Telecommunications Council)

(7) ISAD(G) (General International Standard Archival Description)

(8) MIX (NISO Metadata for Images in XML)

(9) MPEG-7 (Moving Pictures Expert Group)

(10) SEPIADES (SEPIA Data Element Set)

(11) VideoMD

(12) AudioMD

(13) VRA Core

## 3. ISSUES AND CHALLENGES

### 3.1 Creation and Management of Metadata

Librarians and indexers[6] are the specialists who usually create metadata, but recently the non-specialists such as users and publishers are also generating metadata which sometimes is not able to match the standards required for information retrieval. Some wizards and IT tools can be made available in such cases so that the non-specialists enter metadata in a right way. Many word processors and text editors automatically perform the document analysis, filters, extract and generate metadata values themselves and embed metadata tags when the document is created for the first time and metadata is updated when modifications or conversions take place. The guidelines on how, who and when, to write or remove metadata, has to be mentioned clearly in the form of declarative rules and prescriptions. If not the case then the quality and effectiveness of such metadata becomes questionable and such tools have to be included and used extensively in heterogeneous environments to get consensus.

### 3.2 Identifying Core Metadata Elements-Types of Metadata

The primary challenge is to identify fundamental elements needed to describe a resource which are termed as the core elements. Such identified core elements should describe the origin, composition and navigation, confidentiality, legitimacy and recognition, quality and longevity, terms and conditions, of digital objects.

#### *3.2.1 Administrative Metadata*

The information pertaining to the management, organisation, provision and method of access is described for digital collections in the form of administrative metadata. It also provides information about preservation, legal rights, ownership and usage restrictions. METS Rights and MPEG's Rights Data Dictionary[7] (RDD) are excellent schemas for describing intellectual property rights of a digital object in detail. reservation metadata provides information pertaining to provenance, authenticity, preservation activities, and other technical details. The PREMIS schema stands for PREservation Metadata: Implementation Strategies, is a data dictionary for metadata elements which supports XML schema in different contexts, has been designed and developed for sustainable and long term preservation of digital objects.

#### *3.2.2 Structural Metadata*

Structural metadata deals with information relating to the internal organisation of digital resources. For example chapters in a book - indicates how compound objects are put together, i.e., how pages are ordered to form chapters. Structural metadata provides structural relation information with other parent or family files and how the metadata relates to the file, thereby facilitates navigation and presentation of electronic resources.

For example, the audio files created in analog media such as magnetic tape are contextually related to each other and has to be examined while providing structural metadata. The audio files are recorded in different tracks and takes on a tape hence information about other tracks and takes has to be referred in the metadata along with the resources. The metadata about time based media files should also include accurate information about the location points and markers to identify and locate the significant content. In case of individual files within multi-track arrangements, the exact pan location should be included in the structural metadata. In case of a digital excerpt of a larger file the metadata about the resources should be shown in as broader context as possible like an ordered family tree.

#### *3.2.3 Technical Metadata*

The technical qualities of a digital object are described in technical metadata. The technical information such as channel number, bit-depth, sampling rate, frames per second, unique file identifier, etc. AudioMD and VideoMD are an XML-based schema and are extension schemas of METS developed by LOC. A group within the Audio Engineering Society is working on digital library and archive system issues; specifically, on audio preservation and restore. The audio metadata standard they are developing is called AES-X098[8]. The work of the group has parts: (a) descriptive metadata and (b) for administrative metadata to document what are called 'process history' and 'audio object'

#### *3.2.4 Descriptive Metadata*

Descriptive metadata includes curatorial information such as name of the file, creator, associated dates, description, summary, locations etc., is the primary source for the users to access. This provides content-and

context-related specific information of an information resource useful in its discovery and identification. Content metadata includes data about the content of an information resource such as subject terms and titles, Context metadata deals with the data related to the context of an information resource, or data about the creation of that information source. Either the elements in Simple DC[9,10] metadata set can be used directly or an alternative is the MODS for providing descriptive metadata.

There are two descriptive metadata models: (i) Separate metadata model and (ii) Embedded metadata model. In the separate metadata model the metadata is stored in a database linked to resource via some type of locating descriptive metadata such as URL as a separate HTML document linked to the resource it describes. In the embedded metadata model, the descriptive metadata is taken as part of the information they consist of. The metadata is embedded in the web page by the creator using meta tags in the HTML coding of the page, i.e., between the <HEAD> and </HEAD> tags.

### 3.3 Metadata for Institutional Repositories and Information Services

Metadata also exists in various formats such as market reviews, ratings bureaus, critics, discussion forums, blogs, etc. Access to metadata from institutional repositories is either restricted or limited because of various reasons. Research and designing of effective mechanisms for such types of repositories is required. In case of services, the data is dynamic such as weather forecasts which cannot be accessed by an external web harvester. Describing particular domains of resources such as intellectual property restrictions, geospatial coordinates, geostationary satellite imagery, aerial photographs, etc., require specific models to be developed by associating general models of metadata such as RDF. This practical and flexible association supports the amalgamation of varied resources into coherent collections. For example, the Federal Geographic Data Committee Content Standard[11] describes specialised aspects of geospatial data, while a more general model is used to associate that data with related non-geospatial resources.

### 3.4 Metadata for Complex Digital Resources

There are well-established conceptual models and procedures for describing the contents and structure of traditional resources such as text documents. Whereas in case of collections of documents, time-based media, and dynamically generated objects, describing the intellectual contents and structure of resources where variety of encodings are existing, is less understood and far more complex, however there are few conventions in general use. Library of Congress has evolved audioMD and videoMD[12] schemas for time-based media such as audio and video which need 'mechanical' metadata for controlling processes such as synchronisation (e.g., sampling rates and frame sizes), and linked to higher-level descriptive and mechanical metadata through abstractions.

The audioMD[12] schema allows for detailing properties such as: (i) physical format for media-dependent materials (e.g., open reel tape, DAT, instantaneous disc) (ii) track format for magnetic tape recordings (iii) groove type for analog discs (iv) speed and speed adjustment for analog materials (v) number of channels and sound channel map (vi) sampling frequency and bits per sample for digital recordings (vii) audio data encoding for digital materials and (viii) file format name and version for media independent digital materials.

The videoMD[12] schema allows for detailing properties such as (i) physical format for media dependent materials (e.g., stock brand, base and binder for magnetic tapes, disc surface type for disc media) (ii) dimensions for media dependent materials (e.g., diameter, gauge, height, thickness, width, length) (iii) generation for analog media (iv) frame description (e.g., pixel or line counts horizontal and vertical, frame rate) (v) data rate (may be expressed in terms of maximum, minimum, nominal, mode) (vi) format in terms of broadcast standards (e.g., NTSC, PAL, SECAM) (vii) formatting description for digital materials (e.g., name of creating application, commercial name of format, profile, version) and (viii) bits per sample and sampling description in terms of chrominance and luminance (e.g., 4:2:2, 4:2:0).

The metadata for the files store in MPEG-7 format includes information regarding low level descriptions and high level descriptions. It includes various basic, spectral, signal and temporal parameters and information pertaining to terms and conditions, ratings, encoding formats, and scenarios of how multimedia components are combined in presentations. A multimedia content description interface (MPEG-7)[13] technology covers the most recent developments in multimedia search and retrieval, designed to standardise the description of multimedia and related formats content supporting a wide range of applications including DVD, CD and HDTV.

The publications produced in emerging kinds of digital media needs metadata in several aspects such as encoding format, publication type, access controls, description, terms and conditions of use, supporting versions, etc. The metadata information in case of collections of images and manuscript archives has to be different with that of the items in that collection which further help in navigation and discovering specific databases.

### 3.5 Simplicity, Complexity, and Interoperability

The efforts for research in the area of mixing and matching separate metadata schemas, local refinements, specifying interrelations among the modules for simplifying the things in distributed computing environment is necessary if consistency is to be achieved and reduces issues in interoperability. Search engines extract metadata using statistical and algorithmic methods rather than deriving intellectually using semantic analysis methods. There should be a clear understanding of models and strike balance between the simplicity, interoperability, and semantic richness in the metadata schemas.

For example, in case of e-commerce the stuff is complex, metadata is modular and is made up of connecting pieces created by different people. INDECS (INteroperability of Data in E-Commerce Systems)[14], is a metadata initiative for supporting global commerce in intellectual property. Functional Requirements for Bibliographic Records[15] of the International Federation of Library Associations and Institutions (IFLA)[16], describes a range of possible 'states' for information resources from the abstract work through the physical item. PICS (Platform for Internet Content Selection) and P3P (Platform for Privacy Preferences) are the standards intended to provide information on privacy and rating services. The challenge ahead is to define a logical framework[17] which subsumes or reconciles a variety of data models with implications for the switching and reuse of variety of metadata types for a broad range of applications.

### 3.6 Interoperability and Resource Discovery Mapping, Crosswalking and Registries

Interoperability[18] is achieved among networked services and heterogeneous metadata when common conventions on semantics and syntax are agreed and adopted. Sharing common set of elements and modifying them to local requirements across multiple schemas achieves semantic interoperability, whereas syntactic interoperability is possible by strict adherence to standard metadata formats and protocols while creating metadata structures and possible only through metadata deployment. The underlying encoding syntax in RDF is XML which supports exchange of metadata on the web.

Mappings are the relationships between the elements of two metadata standards support access and information transfer across domains. The schemas have significantly different syntaxes or semantics. The metadata element in one schema is linked / mapped with metadata element in another schema by choosing a direct element if exists or a near element if not exists. This technique is extremely useful when mapping old legacy data elements with the new management system or data. UKOLN, an agent for knowledge transfer contains list of existing mappings between many popular schemas which is enormously used. Crosswalks[19] are more complex frameworks that establish the relationship between schemas based on thesauri or ontologies.

Semantic interoperability depends on accurate association of data elements with semantically-related terms within and among systems. This is possible with semantics register[20] in which the data elements and attributes of documents along with meaning and certain other assertions are maintained.

In such registries the elements of different metadata schemas are entered in machine readable syntax along with the authoritative listings of their legal values, local extensions and modifications, mappings to other schemas, and also guidelines and procedures on how to use them. The structured content in these registries perform like knowledge bases which is both human and machine understandable, persuade and endorse the use of such standards formats. Registries[21] support the translation of metadata into other languages and schemas, validation, and updating of remote metadata to new versions automatically. Constructions used in alliance of thesauri, such as 'interlinguas', could be personalised for relating ontologies of metadata.

The Common Object Request Broker Architecture (CORBA) is a standard defined by the Object Management Group (OMG) designed to facilitate the communication of systems that are deployed on diverse platforms. This enables applications to manipulate distributed computing objects; DOI[22], and the Z39.50 protocol for information retrieval.

### 3.7 Information Policies and Management

There are international implications in moving content and metadata between communities and across international boundaries. The information managers should take into account of international implications especially when metadata crosses borders and jurisdictions.

### 3.8 Integrity, Accuracy, and Authenticity of Metadata

A few content providers are submitting irrelevant index terms to the web indexing services[23,24] to get more visibility for their content on the search. This has led to apprehensions and assertions about the content on the network making it increasingly necessary to verify its correctness, reliability, and trustworthiness. Removing misapprehensions and assertions on metadata particularly in case of ratings or subject analysis is a big challenge and the solution is to design and develop algorithms and heuristics. However sampling and analysis

of metadata usage patterns serve as a guide for developers of schemas and provide feedback into the design of core metadata sets and purposeful packages.

## 4. CONCLUSIONS

There is no unique metadata standard sufficient to describe all the documents emerging in various kinds of formats, helpful in efficient information retrieval. Intense and long term research in this area is necessary and possible if metadata specialists and user communities are involved, negotiated, build consensus and try to solve functional problems, particularly in the areas of data organisation, archiving and preservation, rights management and resource discovery of growing networked information sources. Registries are also an important area of research. There is an urgent need to find one unique metadata standard to provide the emerging global information infrastructure with coherent methods of organisation and access that transcend the historical boundaries of nations, languages, and cultures.

## REFERENCES


1. Natasa Sofou, V.T. Metadata descriptors identification and definition-best practice network. 2011. http://www.ECLAP.eu (accessed on 25 August 2014)
2. Higgins, S. What are metadata standards. 2007. http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards. (accessed on 2 September 2014)
3. National Information Standards Organisation Understanding metadata. 2004. http://www.niso.org/publications/press/Understanding Metadata.pdf (accessed on 25 August 2014)
4. Association for Library Collections & Technical Services. Committee on Cataloguing: Description and Access: Taskforce on Metadata. http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html (accessed on 13 September 2014)
5. Zeng, L.M. Metadata interoperability and standardization–A study of methodology. Part I. *D-Lib Magazine,* **2006**. http://www.dlib.org/dlib/june06/chan/06chan.html)
6. ERCIM News Online Edition. Advanced databases and metadata. October 1998. http://www.ercim.eu/publication/Ercim_News/enw35/en35contents.html (accessed on 25 August 2014)
7. B.S. Manjunath. Introduction to MPEG-7: Multimedia content description interface. http://as.wiley.com/WileyCDA/WileyTitle/productCD-04714866787.html
8. Harvard University. Guidance for management and preservation of e-records. 2010. http://isites.harvard.edu/icb/icb.do?keyword=k13588&pageid=icb.page72649 (accessed on 4 September 2014)
9. Baker, T. Languages for Dublin Core. *D-Lib Magazine*. **December 1998**. http://www.dlib.org/dlib/december98/12baker.html (accessed on 8 September 2014)
10. Dublin Core Metadata Initiative. http://purl.org/DC/.
11. ERCIM News Online Edition. Advanced databases and metadata. October 1998 http://www.ercim.eu/publication/Ercim_News/enw35/en35contents.html (accessed on 25 August 2014)
12. Library of congress. Technical Metadata for audio and video. 2014. http://www.loc.gov/standards/amdvmd/ (accessed on 8 August 2014)
13. National Institute of Standards and Technology. MPEG (Moving Picture Experts Group) home page. 2014. http://www.cselt.it/mpeg/
14. INDECS (Interoperability of Data in E-Commerce Systems). 2014. http://www.indecs.org/index.htm (accessed on 10 September 2014)
15. IFLA International Federation of Library Associations and Institutions. Digital librairs: Metadata issues. Evolving Directory of Resources. http://www.ifla.org/ifla/II/metadata.htm (accessed on 8 August 2014)
16. IFLA Study Group on the Functional Requirements for Bibliographic Records. 1998. Functional Requirements for Bibliographic Records. UBCIM Publications, New Series, Vol. 19. http://www.ifla.org/VII/s13/frbr/frbr.pdf
17. Kahn, R.R. A framework for distributed object services. Corporation for National Research Initiatives. May 1995. http://www.cnri.reston.va.us/home/cstr/arch/k-w.html (accessed on 6 July 2014)
18. Bearman, D.E. A common model to support interoperable metadata. *D-Lib Magazine*, **January 1999**. http://www.dlib.org/dlib/january99/bearman/01bearman.html (accessed on 8 September 2014)
19. Lynch, C.A. CNI white paper on networked informatino discovery and retrieval. Coalition for Networked Information. 2014. http://www.cni.org/projects/nidr/www/toc.html (accessed on 7 September 2014)
20. Global Information Locator Services. Making it easier to find all the information. Semantic Registers. 2014. http://www.gils.net/semantics.html (accessed on 4 September 2014)
21. US Dept. of Energy, Berkeley National Laboratory.

Joint Workshop on Metadata Registries - Workshop report. July 1997. http://www.lbl.gov/~olken/EPA/Workshop/report.html (accessed on 6 September 2014)

22. Digital Object Identifier System. http://www.doi.org (accessed on 7 August 2014)
23. Lynch, C.A. CNI white paper on networked informatino discovery and retrieval. Coalition for Networked Information. 2014. http://www.cni.org/projects/nidr/www/toc.html (accessed on 7 September 2014).
24. Lynch, C. Identifiers and their role in networked information applications. ARL, 194. 1997. http://www.arl.org/newsltr/194/identifier.html (accessed in 2014).

**About the Authors**

**Mr P. Ramesh** is presently working as Senior Technical Officer in Dept. of Geo-Physics, University College of Science, Osmania University. He holds MSc (Mathematics), MLISc and UGC NET. He has more than 3 years of teaching experience in the field of library softwares, content management systems and digitisation and development of digital libraries. His areas of interest include: Software development/automation of client/server and distributed applications development with VB and Oracle.

**Mr J. Vivekavardhan** is presently working as Asst. Professor, Dept. of Library Science, University College of Arts and Social Sciences, Osmania University, Hyderabad, Telangana. He has teaching experience of 15 years and has published 25 papers in various journals, and conferences.

**Dr K. Bharathi** is presently working as Asst. Professor, Department of Library and Information Science, Osmania University, Hyderabad. She has published six papers and attended various national and international conferences.