Practical Framework for Harvesting Standard Metadata in Digital Repository

Jiban K Pal

Abstract

Metadata research drastically improved the resource discovery mechanism in accessing information from a large distributed environment. Even technological capabilities permit multiple metadata schemas for standardizing the structure and content of indexing information towards an efficient resource discovery. This paper presents the issues on standard metadata in order to pursue digital repositories and dynamic web pages. It proposes various means of harvesting metadata using optimum standards and protocols. It also enumerates inherent mechanisms for metadata harvesting in DSpace enabled repositories through various harvesting tools; and evaluates XML as current popular choice for metadata harvesting (OAI-PMH) and exchange. System support to multiple metadata formats in DSpace has been discussed thoroughly. Finally it recognizes interoperability and extensibility functions that are being realized increasingly towards a long-term management and preservation of digital objects. However the glimpses of metadata production tools (newer & developed) could accelerate the digital repository initiatives with an increasing popularity of open access movement in real world.

Keywords: Resource Discovery, Metadata Harvesting, Metadata Protocols, Digital Repository, DSpace

1. Introduction

Twenty first century is witnessed an unprecedented change. Various changes in approaches to organizing information and terminologies have been noticed, which occurred from catalogue card to bibliographic record, from bibliographic record to OPAC subsequently metadata, library portal, and library gateway… Metadata is here to stay and evolve, as it becomes a feasible strategy to enhance the resource discovery in a distributed network environment. Functional description of the cataloguers has also been changed. Working cataloguers are increasingly called upon to contribute to digitization projects by creating metadata for digital libraries, harvesting metadata for institutional repositories, selecting metadata standards, defining local application guidelines, and many others. So the concept of metadata has become a buzzword in modern information society. It is equally important for librarians, resource authors, digital archivists, database developers, and seekers of electronic information. In reverse metadata is inevitable for searching i.e. it enables matching of query terms with the terms embedded in the source contents. Particularly, "metadata is expected to improve matching by standardizing the structure and content of indexing or cataloguing information"[1]. Here this paper attempts to present

7th International CALIBER-2009,
Pondicherry University, Puducherry, February 25-27, 2009
© INFLIBNET Centre, Ahmedabad

the issues on creating metadata in order to pursue the digital repositories and dynamic web pages. It also proposes various means of harvesting metadata and enumerates the inherent mechanism for harvesting metadata in DSpace enabled repositories. Discussion brings out the techniques of extensibility and interoperability for presenting metadata using various harvesting tools. Certainly it would be useful to organize an institutional repository with the increasing popularity of open access movement in the publishing world.

2. Memorizing the Concept

The classic definition of metadata is data about data. It describes the attributes and contents of an original document. If an electronic document (read as object) has creator, title, date of creation, etc. then all these elements constitute the metadata about the object. Here this definition entails the basic concept but is perhaps not very meaningful. Basically metadata is an Internet-age term for resource discovery that the librarians have put into catalogues. Most commonly it refers to descriptive information about electronic objects or resources[2]. The term 'metadata' has an ambiguity and difficult to make an explicit definition, but generally refers to – structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource[3]. Many researchers agree that metadata creation is a steady mechanism to maximize the resource discovery in digital environment. Say for example a library catalogue is a collection of metadata elements linked to library documents via call number, information stored in the META field of an HTML page is metadata associated with the information resource embedded within it, indexing data held by web crawlers is also metadata (though not very good metadata) hyper linked to the information resource through the URL[4].

Functions of metadata define its' popular categories and use, say for instance descriptive, administrative, structural, preservation and many other types of metadata varies in their functions[5]. But the prime function of metadata is to help in resource management towards an efficient retrieval in a large digital collection. However it promises rights management, links to e-resources, enables interoperability using standard schemas and protocols, digital object identification (DOI), and so many to facilitates digital preservation. Metadata is an essential phenomenon for online catalogues, federated searching, and open URL's. In fact any digital preservation strategy depends to some extent upon the creation, capture, and maintenance of appropriate metadata. Therefore it is essential for long-term management of digital archives.

3. Recognizing Multiple Metadata Standards

Widespread interest among different metadata communities results the growth of conflicting standards and projects for associating diverse types of metadata with diverse electronic resources. So 'metadata can take variety of forms, may be specialized or general, new metadata sets will develop as network infrastructure matures, different metadata groups will design and be responsible for different types of metadata'[66] .Therefore multiple metadata standards for numerous metadata types can be traced in a hierarchy of complexity. Jan Smits studied the need for various levels of metadata and summarized as – 'if anyone like to describe the complex GIS datasets would probably need to work with FGDC/ISO metadata… MARC can be used with less complex datasets… whereas DC as well as MARC is suitable for raster images and simple vector data sets that do not require a lot of description'. Moreover the demand for uniformity and linkage persists within metadata standards. Suppose the map librarians generally like to create a link between FGDC and MARC or FGDC and DC, minimizing the data entry efforts for OPAC. The inherent cause to keep the records in different formats is basically to enable the interchange of information. Frequently librarians are needed for switching metadata available on their hands into their required standard/s. Virtually the mapping or c**rosswalk** among the standards becomes popular in real practice and such crosswalks within various metadata standards founds available from UKOLN[7]. In view of the above facts a dozen of standards exists for each conceivable digital objects like ETD, e-learning, e-governance, geo-spatial data, museum items, architectural drawings, etc. Such metadata standards include Dublin Core, Meta tags, RDF, TEI, CIMI, GLIS, METS, MODS, MARC, VRA Core, SCROM, LOM, GEM, EAD, PB Core, IMRC, CDWA, CSDGM / FGDC, MIDAS, VERS, DDI, PREMIS, CIDOC, ETDMS, AGLS, ONIX, and so many. Among these standards Dublin Core and Meta tags are widely implemented schemes for describing the content of web resources. Although DC is more widely accepted and used in general, while MARC is popular in the research sector[8]. Dempsey and Heery (2000) divided all metadata standards into three bands – first band includes full-text indexes (eg. search engines as Google), second band emerged to support search and directory services like Spires, Whois++ and even DC too, band three has more complex metadata structures like TEI, MARC, GILS, EAD, etc. Nonetheless every standard has its own specialty. In view of a clear understanding a popular metadata standard has taken under discussion.

4. Dublin Core

Primarily it was developed as a small set of descriptors to describe web based information resources. But quickly it drew global interest from a variety of information providers as an effective tool to discover as well as integrate access to diverse information resources across multiple domains[9]. Actually it was initiated by OCLC through DCMI began in 1995 with an invitational workshop in Dublin (Ohio), to enable more intelligent information discovery systems[10]. Dublin Core Metadata Initiative (DCMI) is an organization dedicated to promote widespread adoption of interoperable metadata standards and specialized metadata vocabularies for describing electronic resources. In fact this standard became finalized in 1996 and defined fifteen metadata elements for resource description in a cross-disciplinary information environment. Such elements are – title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights[11]. These are unqualified DC elements having fifteen core descriptors; but qualified DC has about sixty-five elements and gradually increased over time. A detailed description of the elements founds available at DCMI website. These Core elements can be categorized into three groups on the basis of the type/ scope of information stored in them[12]. These are Content elements related mainly to the content of the resource – Intellectual Property elements related mainly to the resource when viewed as intellectual property – and Instantiation elements related mainly to the instantiation of the resource like date, type, format, and identifier. In 2000, DC got the formal recognition by the Centre for European Normalization (CEN), a european standardization body. Again in 2001, it was ratified under the auspices of NISO and DCMI as ANSI standard (Z39.85-2001)[13]. DC is highly useful in web pages usually stored as name-value pairs within meta-tags and is easy to include at the head of

HTML documents. An example of encoding DC elements using Meta-tags is shown below. Increasingly there are many digital archives of physical objects that are starting to make use of the DC. However, it can also be located in an external document or loaded into a database enabling it to be indexed and manipulated from within a propriety application. A few of the search engines allow for inclusion of limited metadata at the HEADER part, but this metadata could be useful when it follows the recommended syntax for that particular search engine. Guinchard reports the results of her e-mail survey on who uses DC and why and how it is used[141]

DC elements using meta-tags embedded in HTML

```
<HTML>
<head>
<title>
Framework for harvesting metadata in digital repository</title>
<link rel = “schema.DC”  href = “http://purl.org/DC/elements/1.0/”>
<meta name = “DC.Title”  content = “Framework for harvesting metadata”>
<meta name = “DC.Creator”  content = “Jiban K Pal”>
<meta name = “DC.Subject”  content = “metadata, harvesting digital repository”>
<meta name = “DC.Type”  content = “review article”>
<meta name = “DC.Date”  content = “2009”>
<meta name = “DC.Format”  content = “pdf / html”>
<meta name = “DC.Language”  content = “en”>
<meta name = “DC.Identifier”  content = “www.isical.ac.in/~jiban/hmdr.pdf”>
</head>
<body>Content of the object…</body>
</HTML>
```

5. Creating Quality Metadata for Dynamic Objects

Uncountable stacks of resources are available in diverse electronic format demands for creating metadata with quality consistency. One can ask, how we can create metadata for dynamic collection? Whether it can be generated through automatic or traditional means? Who can create a better quality metadata? It is easily understood that traditional techniques (using human efforts) are highly labor-intensive and limiting when large databases or dynamic pages are involved. So the problems of traditional techniques highly demanded for generating metadata by automatic means, which pose a challenge to traditional one. Practically a number of devices like search engine spiders, web crawlers, HTML & XML editors, etc. produce various types of metadata through automatic means. Such devices can generate fairly accurate metadata for a few specific elements - say for date, language, etc. But these tools failed to produce metadata appropriately when it is more intellectually demanded for certain elements like creator, title of the object, subject, etc. However in automatic means there are no consistent filtering practices to ensure the quality/ credibility of metadata. Otherwise certain structural factors in generating software’s or search engine spiders hamper the production of better quality metadata. Therefore, many systems prefer traditional processing exploiting human-intellectual efforts to generate schema-specific metadata.

Again, who can generate metadata with adequate quality? Metadata professionals and resource authors represent two main classes of metadata creators. Metadata professionals (i.e. cataloguers and indexers) have an intellectual ability achieved through training and experience. Obviously they gained their proficiency in the use of content-value and descriptive standards. Although few researchers have noted problems with inter indexer consistency[15]. Ideally professional metadata creators could ensure the efficiency but they are limited in their availability and they never satisfy the law of parsimony. Certainly these professionals can produce high quality metadata[16]. Notionally resource authors make them solely responsible to create the intellectual content of an object. They might also be involved in creating acceptable quality of metadata. “Yet there is a perception that author-generated metadata will be of poor quality and may actually hamper rather than aid to resource discovery”[17]. Greenberg et al reported a counter logic through his study that resource authors have an ability to create adequate quality of metadata as – “…creators are intimate with their work, they want their work to be discovered and consulted, they know their audience and can thus describe their resources appropriately. These factors support the hypothesis that resource authors can create acceptable metadata when working with the DC as this schema initially designed for resource authors… and in some cases they may be able to create metadata that is of better quality than what a metadata professional can produce”[18]. Considering above discussion one can presume and make own conclusion.

So, the creators (like scholars, painters, artists, etc.) regularly creating metadata for their technical or artistic works in the form of abstract, keyword, etc. to make their object more accessible on the web. They are creating metadata through various means like web-forms, web-templates and posting their objects to repositories. In fact most of the digital repositories or open archives (viz. NDLTD, NEEDS, etc) prefer author-generated metadata. Certainly this practice makes sense to produce a consistent and quality metadata in consideration with the phenomenal increase of web-resources and in terms of the economics of hiring professional metadata creators. In such an orientation resource-author normally creates metadata (either by him or under his supervision) at the time of object creation. Several agencies (e.g. FGDC, EPA, etc.) have taken a dominating role in developing web-based metadata entry forms to generate metadata for their particular object. Sometimes the agencies provide a guideline to web-developers on use of ‘meta tagging for search engines’[19]. In real situation, a good number of initiatives (often voluntarily by libraries or by specialists) have been taken so far to catalogue the web resources. Here the OCLC’s InterCat project may be considered as a landmark[20]. Such initiatives are good sign to motivate the information-organizers towards a prospective future of information management.

6. Implementing Metadata in Digital Repository

Metadata is an essential phenomenon for online catalogues, federated searching, open URL’s, etc. i.e. it is inevitable for implementing any digitization projects, data archiving projects, OAI and above all for digital resource management. Usually metadata is embedded in table of contents for books, in meta-tags of web page headers, ID3 for MP3 objects, and in file properties for office documents. Any digital preservation strategy to some extent depends on appropriate metadata implementation. Implementation proceeds through structured formats for metadata harvesting and exchange, say for instance MARC uses ISO-2709 and for header information HTML/ XHTML is useful. However the extensible markup language (XML) is the current popular choice for implementation of metadata, at least to facilitate metadata harvesting (OAI-PMH) or exchange.

7. Metadata Harvesting Tools and Protocols

Harvesting is basically a technique for extracting metadata by automatic means from individual repositories and gathering it in a central catalog to facilitate search interoperability. Harvested metadata may be attached to an object (i.e. encoded in the header of web document), or may be collected in metadata registry or database. Basically the process involves in creation, capture and expose of metadata using protocols. So, a harvester is a client application that recognizes OAI-PMH requests and is operated by service provider as a means of collecting metadata from repositories or open archives. OAI-PMH refers to open archives initiative protocol for metadata harvesting. It is basically a simple protocol that enables regular gathering and transfer of metadata from one system to another. Its' underlying syntax follows common web standards (like HTTP, XML schemas) so as to fairly easy to implement. In fact, OAI-PMH provides an application independent interoperable framework for harvesting to support two classes of clients like data providers (for exposing metadata) and service providers (for building value-added services). OAI-PMH is becoming more popular with the popularity of open access movement in publishing world. Almost all digital repositories and open archives are introducing OAI-PMH to make their metadata available to search engines and harvesters. Even many digital repositories have some inbuilt mechanism to expose metadata using OAI protocol. Therefore, a number of harvesters have developed in the real practice, of which PKP and OAICat becomes more popular. PKP is an excellent open source metadata harvesting and presentation tool[21] developed by John Willinski of university of British Columbia. This multi platform web based tool can effectively extracts metadata and have an intuitive user interface. Another such interesting tool is Virginia Tech Perl harvester that can promise to insert a module as metadata retrieval and browsing program. Similar other harvesters are OAICat, UIUC Java/ VB harvester; DLESE, myOAI and some of them are less tested. Again metadata can be exposed either by using Z39.50 protocol or OAI-PMH and harvesting may be exhaustive or selective. Selective harvesting allows harvesters to restrict harvest requests to portions of the metadata available from a repository and OAI-PMH also supports selective harvesting.

8. Harvesting Metadata in DSpace

DSpace is popular open-source software available for free to anyone and completely customizable for building digital repositories. It captures, stores, indexes, preserves and enables open access to a variety of digital content including text, images, video, audio, animations, etc. DSpace uses OAI-PMH through OAICat (an open-source product of OCLC) for harvesting metadata and can be easily extendable to multiple metadata schemas by developing java programs. Dspace by default uses qualified DC set (has more than sixty-five elements) for furnishing metadata, and exposes metadata using unqualified DC (has fifteen elements) format for the purpose of OAI-PMH. Its' recent versions (1.2.2 beta onwards) allow users to define their own metadata formats by using XML input-forms, i.e. it allow users to extend to non DC formats by modifying $DSPACE_HOME/config/inputforms.xml. Moreover, one can add new elements directly adding to 'dctypeRegistry' table in PostgreSQL. Here the added elements to be indexed in 'dspace.cfg' file, so that Lucene generates indexes on desired elements. Default display can be changed by modifying 'ItemTag.Java' file. Import/ export really does not matter within the DSpace communities but it demands for interoperability mechanism when anybody requires to import/export across other digital library software. Perhaps future versions will permit more integrated use of specialized metadata. In view of this MIT's SIMILE project is investigating semantic web technologies. No doubt the support for multiple metadata formats may greatly enhance the use of DSpace for archiving the digital objects. Dr Prasad in a user meet at Cambridge has made a detailed discussion[22] in this direction. However, DSpace primarily deals with three types of metadata for the archived content[2323] DSpace Federation at MIT: DSpace system documentation - metadata. Source: http://www.dspace.org/technology/system-docs/functional.html (accessed on 21st October, 2008).**Brief Profile of Author** **Jiban K. Pal** (b.1972) holds B.Sc in Zoology, B.Ed, MLIS and started his carrier with the Indian Statistical Institute Library, Kolkata in 1997. Currently Mr. Pal is working in the Periodicals Unit of Library Documentation & Information Science Division and entirely involved in library automation activities of the same Institute. His area of interest lies in library automation, consortia, scholarly journal pricing, scientometrics and information retrieval. His current interest reflects on information retrieval and open repository. He has several published and communicated papers to his credit and participated in various national & international level workshops and conferences. He is the life member of BLA, IASLIC and ISI. – namely descriptive (for description), administrative (for preservation, authorization policy data, etc.), and structural (for presentation i.e, implementation of METS).

9. Conclusion

Phenomenal growth on metadata research and content organization techniques drastically improved the precision of resource discovery in distributed network environment. Now one can retrieve any digital repository more accurately that have been catalogued using adequate quality metadata. Hardly any digital preservation project can survive without harvesting appropriate metadata. Support for multiple metadata formats in Dspace greatly enhanced for building digital repositories through out the world. Increasingly it has been realized by the information community towards the extensibility and interoperability function of metadata that could bring a reasonable solution towards producing high-quality metadata for digital archiving. Above all the integration of metadata sets together and development of new metadata production tools would be a great frontier in future information science research.

References

1. **Milstead, Jessica & Feldman, Susan** (1999): Metadata – Cataloging by any other name… in *Online*, (© Information Today Inc.) January issue. Source: http://www.infotoday.com/online/OL1999/milstead1.html (accessed on 12th June 2008).
2. **DCMI** – Introduction: What is Metadata? Source: http://dublincore.org/documents/usageguide/ (accessed on 12th June 2008).
3. **NISO:** Dublin Core Metadata Element Set Approved. Source: http://www.niso.org/news/releases/PRDubCr.html (accessed on 14th June 2008).
4. **Cathro, Warwick** (1997): Metadata – an overview. Source: http://www.nla.gov.au/nla/

staffpaper/cathro3.html (accessed on 2nd July 2008).

5. **NISO** (2004): Understanding Metadata. Bethesda, MD: NISO Press, p.1-12

6. **Lagoze, Carl** (1996): The Warwick framework – a container architecture for diverse sets of metadata, in *D-Lib Magazine*, July–August issue. Source: http://www.dlib.org/dlib/july96/lagoze/07lagoze.html (accessed on 12th December 2008).

7. **Michael Day** (2002): Metadata - mapping between metadata formats, May 2002, UKOLN - The UK Office for Library and Information Networking, University of Bath, UK. Source: http://ukoln.ac.uk/metadata/interoperability/ (accessed on 14th July 2008).

8. **Polydoratou, P & Nicholas, David** (2001): Familiarity with and use of metadata formats and metadata registries amongst those working in diverse professional communities within the information sector, in *Aslib Proceedings*, 53 (8), p.309-324.

9. **Quam, Eileen** (2001): Informing and evaluating a metadata initiative – usability and metadata studies in Minnesota's Foundations Project, *Government Information Quarterly*, v.18, p.181-194.

10. **DCMI** – Dublin Core Metadata Initiative. Source: http://dublincore.org/ (accessed on 15th July 2008)

11. **DCMI**: Dublin Core Metadata Element Set, Version 1.1: Reference Description. Source: http://dublincore.org/documents/dces/ (accessed on 14th July 2008).

12. **Burnett, K & Ng, K. B. & Park, S** (1999): A comparison of the two traditions of metadata development, in *Journal of the American Society for Information Science*, 50 (13), p.1209-1216.

13. **Dekkers, Makx & Weibel, S. L** (2002): Dublin Core metadata initiative progress report and workplan for

2002, in *D-Lib Magazine*, 8 (2), p.1-9. Source: www.dlib.org/dlib/february02/weibel/02weibel.html (accessed on 12th August, 2008).

14. **Guinchard, Carolyn** (2002): Dublin Core use in libraries – a survey, in *OCLC Systems & Services*, 18 (1), p.40-50.

15. **Chan, L. M.** (1989): Inter-indexer consistency in subject cataloging, in *Information Technology & Libraries*, 8 (4), p.349-358.

16. **Weinheimer, J.** (2000): How to keep the practice of librarianship relevant in the age of the Internet. vince (special issue on Metadata, part 1), v.116, p.14-27.

17. **Thomas, C & Griffin, L** (1999): Who will create the Metadata for the Internet?, in *First Monday* – a peer reviewed journal of the Internet. Source: www.131.193.153.231/issues/issue3_12/Thomas/index.html (accessed on 18th August, 2008).

18. **Greenberg, J.** & et. al (2001): Author-generated Dublin Core metadata for web resources – a baseline study in an organization, in *Proceedings of the International Conference on Dublin Core and Metadata Application*, held at NII, Tokyo, Japan on October 24-26, 2001, p.38-46

19. **Richmond, Alan**: META tagging for search engines, in *Web Developer's Virtual Library*, Source: http://www.wdvl.com/Search/Meta/Tag.html (accessed on 22nd September, 2008).

20. **OCLC**: InterCAT Project. Source: http://www.oclc.org/research/projects/archive/intercat.htm (accessed on 24th September, 2008).

21. **Kellogg, David** (2004): Open source OAI metadata harvesting tools. Source: http://www.diglib.org/aquifer/oct2504/harvesting.pdf (accessed on 16th October, 2008).

22. **Prasad, A R D** (2005): Using Multiple Metadata formats in DSpace, presented at an User Meet on 6-8th July 2005, University of Cambridge, UK

23. DSpace Federation at MIT: DSpace system documentation - metadata. Source: http://www.dspace.org/technology/system-docs/functional.html (accessed on 21st October, 2008).

About Author

Mr. Jiban K Pal, Periodicals Unit of Library Documentation & Information Science Division.