Annals of Library and Information Studies
Vol. 57, March 2010, pp.44-53

Metadata initiatives and emerging technologies to improve resource discovery

Jiban K. Pal
Library, Documentation & Information Science Division, Indian Statistical Institute,
203, B. T. Road, Kolkata – 700 108, Email: jiban@isical.ac.in

This paper recognizes some emerging issues on metadata as a mechanism of resource discovery and its impact on precision of search results in a distributed network environment. It aims to present a brief account of the major metadata initiatives taken during the last couple of years, thus provide glimpses of recent activities on metadata across the globe. It also highlights a consistent growth of multiple metadata standards to meet the variety of needs in a hierarchy of complexity. The paper examines various metadata-harvesting tools and related technologies that fulfill the task implicit in a user's search. Discussion brings out some popular standards, useful protocols, and open-source harvesters along with their intrinsic capabilities for harvesting and presenting metadata. It also emphasizes on a variety of metadata services viz., OCLC's metalogue service, UKOLN metadata editor service, OAIster harvester service, DP9 gateway service, etc. that are predominantly used in different metadata communities. Attempt has been made to explore the underlying principles of metadata-harvesting in DSpace and web search engines. It also seems imperative to make a discussion on the use of multiple metadata formats in DSpace enabled archives for exposing domain-specific metadata; and subsequently evaluates the inherent mechanism for extensibility and interoperability functions. Thus it proposes various means of creating metadata in order to pursue high-precision document retrieval in dynamic collections. Finally it notices semantic web technologies that could bring a reasonable solution towards the integrated use of specialized metadata for long-term management and preservation of digital objects.

Introduction

Digital resources are growing at an exponential rate and uncountable stacks of resources are available on the web. Semantic based search engines and meta-search engines stimulate resource discovery on digital collections[1]. So the users can get multiple sources that are relevant to their queries. But huge resources practically create a formidable hurdle for accessing desired information effectively and efficiently. In fact, a considerable amount of noise always exists in retrieval of information, which is basically due to uncountable number of heterogeneous resources available in a large distributed environment. In view of this situation, metadata creation is an effective strategy to enhance the resource-discovery from a digital collection. However standard guidelines are essential in creating metadata with quality and consistency that can be accomplished by standard metadata schema. Consequently, standards ensure compatibility and facilitate interchange ability of information sources across the global network system. It also improves quality of information services and reduces economic and technical barriers in information flow[2]. Creation of standard metadata requires extra skills and can be possible either by embedding structured metadata in web-resource headers or through installing a metadata search engine (e.g. HotMeta). In fact technological capabilities now allow multiple metadata schemas in producing metadata from complex digital environment.

Recognizing metadata concept

Metadata today is an essential phenomenon for electronic cataloguing, federated searching, and open URL's. Increasingly, working cataloguers are called upon to contribute to digitization projects or institutional repository projects for creating metadata, selecting metadata standards, identifying metadata harvesting tools, assigning local application guidelines, and many others. So, metadata is perceived to be essential for the librarians in pursuing long-term management and preservation of digital objects. Metadata is also important for digital archivists, database developers, resource authors, web-page designers, aggregators, system designers, as well as seekers of electronic information. In fact, metadata is inevitable for digital resource management and for discovering information from a large distributed environment. Particularly, "metadata is expected to improve matching by standardizing the structure and content of indexing or cataloguing information"[3].

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[4]. The term 'metadata' has an ambiguity and it is difficult to make an explicit definition, but generally it refers to – structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource[5]. 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[6].

Functional diversification and categorization

Various types of metadata have their own functions. Descriptive metadata enumerates the object to discover or identify the information resources, whereas administrative metadata depicts information to administrate and manage the resources that includes legal rights to access (IPR), when and how created, version control, etc. Similarly when structural metadata describes the way of bringing similar resources or compound objects together, then technical metadata indicates the system functions and technical behavior (viz. formats, compression ratio, data authentication, encryption keys, etc), and preservation metadata provides information required for preservation management like archiving the resources, physical features, survival challenges, etc.; and many other types of metadata varies in their functions[7]. So the diversified functions of metadata define its popular categories and use. A significant number of writings however, focus only on the function to support 'resource discovery'. It means the prime function of metadata is to help in resource management towards an efficient retrieval in a large digital collection. Metadata not only supports resource-discovery but also promises rights management, links to e-resources, enables interoperability using standard schemas and protocols (e.g. cross search by Z39.50 protocol or metadata harvesting using OAI protocol), digital object identification (DOI), and digital preservation. In fact metadata can make it possible for users to determine the availability and usefulness of information 4 i.e., whether the information objects exist, how many and where are the objects, whether the objects are useful, authentic, etc. Strebel et al describes three main functions of metadata viz., data access, data management, and data analysis[8]. However metadata functions can also be described in two different levels – one is system level where metadata provides facility for interoperability or integrity of resource discovery tools; another is end-user level where metadata ensures capacity to determine the type of data available, how to acquire it, whether meets the requirement, how to capture at user-end, etc.[9]. In principle, metadata acts as surrogate for a larger whole and makes the resource objects available to end-users – hence it is functionally justified.

Metadata initiatives and global trends

While there are disparate sets of needs to formalize and standardize metadata, several attempts have been made by libraries, federal agencies, voluntary organizations, and others to satisfy the perceived interests of those communities. This widespread interest among different metadata standard groups has resulted in the growth of conflicting standards. Therefore, "metadata takes a variety of forms, both specialized and general — new metadata sets will develop as the networked information infrastructure matures — different communities will propose, design, and be responsible for different types of metadata[10]". However the situation stimulated metadata communities to meet and talk earnestly all over the world through various workshops (Dublin Core workshops), conferences (IEEE and LC conferences on metadata), seminars (OCLC seminars on metadata, offered regularly) and meetings. Even various standard setting bodies, working groups, task forces of different organizations like International Organization for Standardization (ISO), National Information Standards Organization (NISO), Dublin Core Metadata Initiative (DCMI), American National Standards Institute (ANSI), National Committee for Information Technology Standards (NCITS), Federal Geographic Data Committee (FGDC), Library of Congress (LC), Online Computer Library Center (OCLC), UK Office for Library and Information Networking (UKLON), International Federation of Library Associations and Institutions (IFLA), North Carolina State University (NCSU), and many others have emerged.

Several projects have initiated research on metadata which include the DESIRE project[11], OCLC's CORE project, Alexandria project, IEEE's LOM/ Sharable Content Object Reference Model (SCORM) project, Government Information Locator Service (GILS) project, ROADS project, MetaWeb, Nordic Metadata, Computer Interchange of Museum Information (CIMI), Text Encoding Initiative (TEI), Encoded Archival Description (EAD), Content Standard for Digital Geospatial Metadata (CSDGM), and many others[12]. It has to be mentioned that LC undertook the first formal initiative in 1960's on the MARC. In view of its comprehensiveness, interoperability, and maturity - MARC is highly specific and holds semantically enriched metadata. But in 1990's, remarkable growth of digital repositories on the web has been noticed. MARC alone cannot be used for exploring different organizational repositories. Gradually, different professional communities have introduced new ideas, standards, guidelines, and architectures for managing the growing digital resources.

The CSDGM was initiated by Federal Geographic Data Committee (FGDC) in 1992; NCSU libraries introduced EAD in 1993; Dublin Core Metadata Initiative (DCMI) began in 1995. Simultaneously, in mid 1990's – CIMI consortium initiated metadata test-bed project as an extension of DC elements; TEI emerged from humanities and linguistic research communities; GILS developed a complex metadata format with an intention to identify the US government information resources; ROADS project has undertaken to design and implement the user-oriented resource discovery system by UKLON; SCORM uses IEEE's Learning Object Metadata (LOM) element set for descriptive metadata. In this regard, W3 consortium has taken a strong initiative in metadata and its standards are very simple (meta-meta level) with an attention to make it highly compatible to a variety of designs. For the purpose, W3 consortium has developed Resource Description Framework (RDF) and PICS specifications to be used to encode and transmit the metadata produced from DC and Warwick Framework[13].

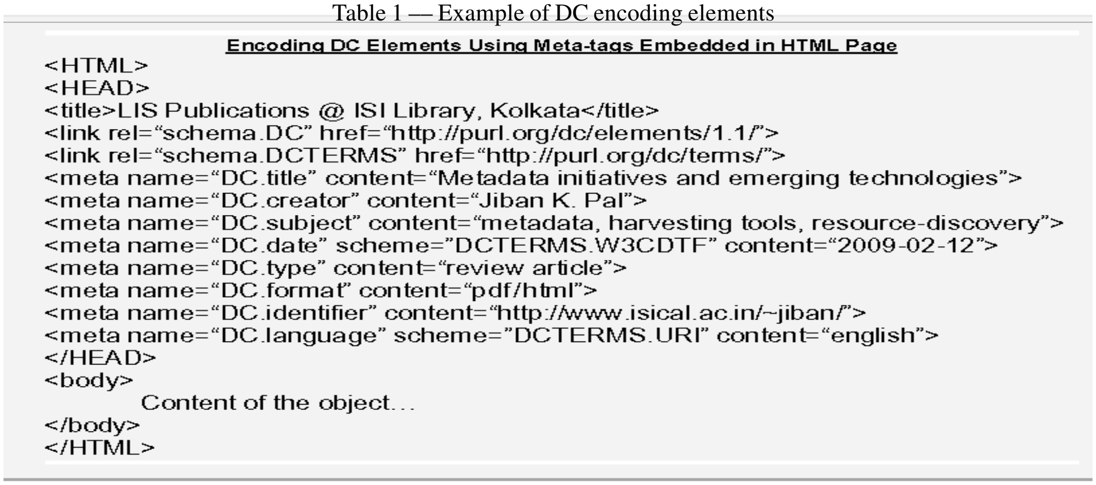
Three patterns of metadata standards evolution are seen, (i) metadata standards that evolved in different professional traditions (MARC, CSDGM, CDWA, etc.), (ii) metadata standards that evolved in flexibility and scalability in metadata structures (Dublin Core, etc.), and (iii) metadata standards that evolved in adoption of a common formal language to support different applications (SGML, XML in application with EAD and TEI, etc.). Although some existing ideas and projects are in progress and some concerns about the future directions in organizing web resources still remain. A recent survey entails current projects and initiatives on metadata research for organizing digital resources[14].

Growth of multiple metadata standards

Hence a single standard cannot suit all; therefore multiple metadata standards for numerous metadata types can be traced in a hierarchy of complexity. Jan Smits[15] studied the need for various levels of metadata and summarized as – 'anyone who likes 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 faster 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 like to create a link between FGDC and MARC or FGDC and DC, thus 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). Thus mapping or crosswalk among metadata formats has become a popular practice in modern libraries, which is available from UKOLN[16] site.

In view of the above facts, a dozen of standards exist 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, e-GMS, ONIX, and many others. Among these standards Dublin Core and Meta tags are widely

Table 1 — Example of DC encoding elements



Encoding DC Elements Using Meta-tags Embedded in HTML Page

```
<HTML>
<HEAD>
<title>LIS Publications @ ISI Library, Kolkata</title>
<link rel=“schema.DC” href=“http://purl.org/dc/elements/1.1/”>
<link rel=“schema.DCTERMS” href=“http://purl.org/dc/terms/”>
<meta name=“DC.title” content=“Metadata initiatives and emerging technologies”>
<meta name=“DC.creator” content=“Jiban K. Pal”>
<meta name=“DC.subject” content=“metadata, harvesting tools, resource-discovery”>
<meta name=“DC.date” scheme=”DCTERMS.W3CDTF” content=“2009-02-12”>
<meta name=“DC.type” content=“review article”>
<meta name=“DC.format” content=“pdf/html”>
<meta name=“DC.identifier” content=“http://www.isical.ac.in/~jiban/”>
<meta name=“DC.language” scheme=”DCTERMS.URI” content=“english”>
</HEAD>
<body>
        Content of the object...
</body>
</HTML>
```

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[17]. While PREMIS (by OCLC) has been developed specifically to support digital preservation, e-GMS (e-government metadata standard) is popular for information resources across government and public sector in UK. Dempsey and Heery[18] devised an approximate typology of semantically richer metadata formats based on their shared characteristics such as method of creation, search and retrieve protocols, status, etc. They placed all metadata standards in three different bands – where first band derived metadata from full-text indexes (eg. search engines as Google); second band to support search and directory services like Spires, WHOIS++ and even DC too; band three for more complex structured metadata formats (viz. FGDC, MARC, GILS) or having a larger semantic frameworks (viz. TEI, EAD, CIMI) usually developed to meet domain-specific requirements. Nonetheless every standard has its own specialty. For a clear understanding, a few popular metadata standards have been discussed here.

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[19]. Actually it was initiated by OCLC though DCMI began in 1995 with an invitational workshop in Dublin (Ohio), to enable more intelligent information discovery systems[20]. 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. This standard was finalized in 1996 with fifteen metadata elements for resource description in a cross-disciplinary information environment. The elements are – *title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage,* and *rights*[21]. 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 is available at the DCMI website. These Core elements can be categorized into three groups on the basis of type or scope of information stored in them[22]. These are ‘content elements’ related mainly to the content of the resource – ‘intellectual property elements’ related 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 formal recognition by the Centre for European Normalization (CEN, a standardization body). Again in 2001, it was ratified under the auspices of NISO and DCMI as ANSI standard (Z39.85 – 2001) [23].

DC is highly useful for encoding metadata in web pages usually stored as name-value pairs within meta-tags and fairly easy to embed into the header of HTML documents. An example of encoding DC elements using meta-tags is shown in Table 1. 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 an e-mail survey on who uses DC and why and how it is used[24].

Federal Geographic Data Committee (FGDC)

FGDC which is a nineteen member inter-agency committee in US, has established this standard for digital geospatial objects. This nationwide data publishing effort, supported by National Spatial Data Infrastructure (NSDI), is primarily intended to coordinated sharing and dissemination of geospatial data. As such federal government imposed a mandate for the federal agencies to generate FGDC based metadata for their digital objects and many other agencies also use this standard. In fact FGDC metadata has received its widespread acceptance as FGDC records can be indexed more extensively than MARC. No doubt this standard is inevitable for many agencies like Alexandria Project or Harvard's Laboratory. However, this standard is not sophisticated enough to suit everyone's needs and is often criticized as it is too cumbersome and more difficult to apply. Even FGDC metadata is featured by its incompatibility with MARC. Therefore a number of initiatives are underway to harmonize FGDC standard with European standards. Perhaps its' new fragment might supercede the current one viz., CSDGM, FGDC/ ISO, FGDC framework, etc. However many agencies/ states in US are eagerly trying to adopt a simple version of FGDC metadata (called as metadata-lite) that concentrates on essential elements for providing better flexibility in data translation. It is worthy to mention that Content Standard for Digital Geospatial Metadata (CSDGM) is an extension of FGDC and emerged to provide a common set of terminology and definitions for the documentation of digital geospatial data.

MARC/ MARC21

It is an acronym that stands for MAchine Readable Cataloguing. It was primarily designed to serve the needs of libraries as a convenient way of storing and exchanging bibliographic information. Library of Congress designed original MARC format in 1965–66 leading to a pilot project known as MARC-1. Gradually several versions like USMARC, UKMARC, CANMARC, INTERMARC and as many as twenty formats have emerged. Since these formats differ from each other and owing to the problem of inconsistency among different MARC formats, the UNIMARC was developed to accommodate records created in any of the MARC format. Thus the records from one MARC format could be converted into UNIMARC and then UNIMARC to another. The metamorphic phases of MARC led to MARC21 which is the first harmonized version of USMARC and CANMARC. It established a common taxonomy for defining the contents of print and electronic documents. Johnson remarks that despite the great potential of XML, MARC continues to be important and is a well accepted encoding system[25]. Now MARC21 has great potential to describe computer-readable bibliographic records and majority of library automation systems allow for data entry, indexing, retrieval, and display in this format, even if the records are stored internally in another format.

Z39.50 protocol can be useful to execute searches of MARC fields from a Z39.50 client to a Z39.50 server fronting a database of MARC records, and retrieved records can be returned in MARC format[26]. MARC21 format is a set of codes and content designators defined for encoding machine-readable records. Basically this format is defined for five types of data such as bibliographic, holdings, authority, classification, and community information. However, a MARC record involves three elements, namely record-structure, content-designation, and data-content[27]. Record-structure consists of three main components like leader, directory, and variable fields. Content-designation objectively identifies and characterizes the data elements with sufficient precision to support manipulation of data for a variety of functions like display, retrieval, etc. Data Content implies the content of most data elements is defined by standards outside the formats like *AACR, LCSH* but the content of other data elements (i.e., coded data) is defined by MARC21 formats.

Metadata harvesting tools and services

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 creation, capture and expose of metadata using preset standard and protocols. Thus 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. So, typically the metadata harvesting requires three basic tools (at least) viz. the protocol, the harvester, and the standard or schema. Usually two kinds of protocols become useful for enabling metadata in distributed network environment; these are OAI-PMH and Z39.50 or Zing. 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) that are fairly easy to implement. In fact, OAI-PMH provides an application independent interoperable framework and supports Dublin Core elements. This protocol uses six verbs to perform various functions; such as a harvester uses these verbs to harvest metadata across digital repositories. OAI-PMH is becoming more popular with the popularity of open-access movement in publishing world. It facilities both classes of clients like data providers (for exposing metadata) and service providers (for building value-added services). This protocol has been supported by a good number of digital repositories to make their metadata available to harvesters and search engines. Even many digital achieves have some inbuilt mechanism to expose metadata using this protocol, such as Librarians' Digital Library (LDL), Search Digital Libraries (SDL) of Documentation Research and Training Centre (DRTC) in India. However, Z39.50 is traditionally used protocol, developed primarily to search OPACs in the library parlance. Over time this protocol evolved with expanded functionalities to perform real-time retrieval from digital repositories. But, convergence of Z39.50 with web-technologies (i.e. WWW, SRU, SRW, HTTP, SOAP, etc) lead to emergence of an enhanced protocol called ZING (Z39.50 International for Next Generation). Both the protocols have their intrinsic capabilities for exposing metadata, but Z39.50/ Zing is less used and needs further improvement in this particular context, whereas OAI-PMH is more extensively used in real practice. Notably, metadata harvesting may be exhaustive or selective and OAI-PMH also supports selective harvesting.

A good number of harvesters are available in real practice but their usefulness varies with the functional requirements in harvesting of demanded metadata. Therefore choice for a simple and appropriate harvester is obvious. Among the open source harvesters PKP, UM Harvester, OAICat, and Virginia Tech Perl proved to be popular harvesting tools. PKP harvester is an excellent open source metadata harvesting and presentation tool developed by Public Knowledge Project (John Willinski) from University of British Columbia[28]. This command-line software fits well with scheduling tasks and can be installed easily into a LAMP-based server (Linux-Apache-MySQL-PHP) without writing any configuration file. This multi-platform web based tool has adequate flexibility in harvesting OAI metadata in a variety of schemas, including unqualified DC, MODS, MARCXML. PKP open archive harvester uses an impressive GUI and has an intuitive user interface. University of Michigan's UMHarvester is a simple metadata harvesting tool that places harvested metadata into directories and is very easy to browse. Similarly OAICat harvester is an effective open-source tool developed by OCLC. It provides a repository framework that conforms to OAI-PMH and can be customized to work with multiple metadata schemas in performing arbitrary operations on data harvesting by implementing some Java interfaces. Another such interesting tool is Virginia Tech's Perl harvester. This command-line harvester is very flexible and promises to insert a module for metadata retrieval. Some other harvesters and harvesting services are also available in public domain. A few of them are less tested and least used, these are UIUC-Java harvester, UIUC-VB harvester, myOAI, ODL, Ivia, CPAN OAI harvester, etc. However, several gateway services provide intermediation for many digital repositories by harvesting their metadata, such as OAI static repository gateway, Emory's metadata migrator, UIUC's FileMakerPro, Z39.50 gateways, etc. Furthermore, DP9 (an OAI gateway service for web crawlers) at Old Dominion University is a harvester that enables search engines to harvest records from OAI-PMH repositories; DRTC's SDL harvester currently indexed 39053 objects from 24 digital archives; OAIster harvester service (collaborative product of OCLC and Michigan Univ.) currently provides access to 20,220,634 records from 1082 repositories; ZMARCO allows MARC records which are already available through Z39.50 server to relatively easily be made available via the OAI-PMH. Finally, the standard or schema forms a guideline to harvest objective-specific metadata with preset criteria.

It is worthy to mention that a variety of metadata services are predominant in different domains. UKOLN site provides metadata editor service that can automatically generate DC metadata codes either in HTML or RDF/XML format for embedding in the header of web pages.

Even such generated metadata codes can also be converted into various other formats like IEEE-LOM, USMARC, SOIF, IAFA/ROADS, TEI, GILS, IMS or RDF[29]. Metadata librarians are frequently needed for switching metadata available on their hands into their required standard/s and mapping or crosswalk among the standards found available from UKOLN. However, MIT's metadata service (fee-based) unit provides consultation to other libraries and offers extensive expertise in metadata applications, workflow planning, interoperability assessment, schema design, project evaluation, training, etc. Similar services are also available from OCLC's Metalogue, HWW's TV metadata service (in Australia), Oregon University Library's MSDP service, Cornell University Library's digital consulting and metadata service, Colorado University Library's metadata service, and others. DRTC's DLR group provides solutions to metadata related problems in Indian libraries, especially metadata issues in DSpace enabled repositories – as DRTC is a recognized test-bed center for DSpace in India.

Metadata harvest in DSpace and Web Search-engines

Metadata is essential for digital curation. Usually metadata is embedded in table of contents of books, in meta-tags of web page headers, in ID3 of MP3 objects, and in file properties for office documents. Any digital preservation strategy to some extent depends on appropriate metadata organization that can be possible through structured formats. For instance MARC uses ISO-2709 and HTML/ XHTML is useful for header information. However the extensible markup language (XML) is the current popular choice for implementing metadata, at least to facilitate metadata harvesting or exchange. Here the inherent mechanism of harvesting metadata in DSpace and web search engines is explored.

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. Prasad in a user meet at Cambridge has made a detailed discussion in this direction[30]. However, DSpace primarily deals with three types of metadata for the archived contentt:[31] – namely descriptive (for description), administrative (for preservation, authorization policy data, etc.), and structural (for presentation i.e, implementation of METS).

Web search is undergoing an evolution from high-precision document retrieval for keyword queries, to fulfilling the task implicit in a user's search. Search engines and meta-search engines have pivotal role in discovering the resources on the web. Search engine allow users to search and access the resources from a distributed network for managing information overload, whereas meta-search engine allow users to access many search engines together and retrieve ranked result following global merging techniques[32]. Dozens of search engines (yahoo, google, infoseek, etc.) has greatly expanded the access to any digital information as the collections are enabled using standard metadata. An experiment made by Turner and Brackbill, found that addition of keywords within meta-tags of web pages drastically improve the retrieval ability of many search engines[33]. Although a few do not look for meta-tags in their search techniques like Yahoo, Excite, etc. So, common metadata standards can make search engines more efficient of which meta-tags and Dublin-core has great implementation value. In recent times, commercial publishers and other publishing societies are taking initiatives to make their online content more visible to popular search engines. In fact AIP, IoP, IEEE, OCLC's WorldCat, etc undertook major initiatives to index their digital content through Google[34] for greater visibility. Even the commercial publishers are keen to have a cross-publisher citation linking system that began as a pilot collaboration with Google in 2004 to allow indexing of full-text content from more than 29 academic publishers.

In this regard, Mohamed examined the impact of using metadata in discovering the web resources. His study claims that metadata elements can highly influence the page rank order. Even the rank order of the pages that contain meta-tags is higher than the pages those include DC and those do not contain any metadata[35].

Debates in metadata creation

Growing web resources in diverse electronic formats demand creating metadata with quality and consistency. Lack of adequate quality metadata may result in failure to discover the relevant objects when it is actually needed. Obvious enquiries are, how one can create metadata for dynamic objects? 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 resources are involved. So the problems of traditional techniques highly demand for generating metadata by automatic means, which pose a challenge to traditional one. In fact a number of devices like search engine spiders, web crawlers, HTML and XML editors, etc. produce numerous types of metadata through automatic means. Practically such devices can generate fairly accurate metadata for some specific elements (date, language, etc.); but they failed to produce metadata consistently when it is more intellectually demanded for certain elements like creator, subject, title of the object, etc. However, in automatic means there is no consistent filtering practice to ensure the quality and credibility of extracted metadata. Otherwise some structural factors in generating software's and search engine spiders bring displeasure in producing optimum quality metadata. Therefore, many systems prefer traditional processing so as to generate schema-specific metadata using human-intellectual efforts.

Further, who can create 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[36]. Ideally professional metadata creators could ensure the efficiency but they are limited in their availability and too costly – so as to violate the law of parsimony. Certainly these professionals can produce high quality metadata[37]. 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[38]". Greenberg et al established counter logic and reported 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 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[39]". In practice, 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. Again, 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'[40]. Practically a good number of initiatives (often voluntarily by libraries or specialists) have taken so far to catalogue the web resources and OCLC's InterCat project is considered as landmark[41]. Such initiatives are indicative for a prospective future of information management. So, considering the above discussion information organizers can presume and draw their own conclusions.

Conclusion

Metadata is an essential phenomenon for online catalogues, federated searching, and open URL's. Any digital preservation strategy essentially depends upon the creation, capture and maintenance of appropriate metadata. Recent growth on metadata research and content organization techniques enormously improved the resource discovery mechanisms in accessing information from a distributed network environment. Technological capabilities now allow multiple metadata schemas for standardizing the structure and content of indexing information towards an efficient resource discovery. Further enhancement in discovery-mechanism is also possible; if all digital objects can be catalogued obligatorily with a single metadata format and same controlled vocabularies can be used for producing consistent metadata. Practically it would be very difficult (though not impossible) to define a single metadata format for diverse electronic objects archived in cyberspace – as metadata sets differ in potentialities to meet a variety of needs. However, the support for multiple metadata formats in digital repository software's (DSpace, etc.) greatly enhanced the open-access movement through out the world. Therefore, integration of metadata sets and development of new metadata harvesting tools would be a great frontier in future information science research. Gradually it has been realized by the information community towards interoperability and extensibility functions that could bring a reasonable solution for producing high-quality metadata in digital collections. Publishers, consortium managers, aggregators, database developers, and librarians (especially cataloguers, metadata coordinators) should have some responsibility for re-initiating metadata policies with additional priorities that will be another challenge to make the existing standards potentially useful. Overall, metadata is perceived to be essential for the long-term management and preservation of digital objects.

Acknowledgements

The author gratefully acknowledges and thanks Dr. Usha Mujoo Munshi, former Chief Librarian, Indian Statistical Institute for her inspiration and valuable comments on the draft of this paper. He also wishes to thank all of his beloved colleagues for providing their constant support to prepare this paper.

References

1. Ding W and Marchionini G, A comparative study of web search service performance, In *Proceedings of the American Society for Information Science*, (Baltimore: October 21-24) (1996), p.136-142.
2. Persico J Jr., Leadership and empowerment in a total quality organization, *Total Quality Management*, 2 (1) (1991) 57-67.
3. Milstead J and Feldman S, Metadata: Cataloging by any other name, *Online* (© Information Today Inc.), January issue (1999), Available at: http://www.infotoday.com/online/OL1999/milstead1.html (Accessed on Accessed on 12 May 2008)
4. DCMI, Introduction: What is Metadata, Available at: http://dublincore.org/documents/usageguide/ (Accessed on 12 May 2008)
5. NISO, Dublin Core Metadata Element Set Approved, Available at: http://www.niso.org/news/releases/PRDubCr.html (Accessed on 18 May 2008)
6. Cathro W, Metadata: an overview, (1997), Available at: http://www.nla.gov.au/nla/staffpaper/cathro3.html (Accessed on 22 June 2008)
7. NISO, Understanding Metadata, Online edn (NISO Press; Bethesda, MD), 2004, p.1-12
8. Strebel D, Meeson B and Frithesen J, Metadata standards and concepts for interdisciplinary scientific system – II, Position paper from IEEE Metadata Workshop (Washington D C: May 1994).
9. Bor-Ng K, Park S and Burnett K, Control or management: a comparison of the two approaches for establishing metadata schemes in the digital environment, Available at: http://www.scils.rutgers.edu/~sypark/asis.html (Accessed on 16 July 2008)
10. Lagoze C, The Warwick framework: a container architecture for diverse sets of metadata, *D-Lib Magazine*, July-August issue (1996), Available at: http://www.dlib.org/dlib/july96/lagoze/07lagoze.html (Accessed on 18 July 2008)
11. Dempsey L and Heery R, Specification for resource description methods (Part 1) - A review of metadata: a survey of current resource description formats - some characteristics of investigated metadata formats), In *Work Package 3 of Telematics for Research project DESIRE (RE-1004)*, UKOLN Metadata Group, Available at: http://www.ukoln.ac.uk/metadata/desire/overview/rev_01.htm (Accessed on 10 August 2008)
12. Milstead J and Feldman S, Metadata projects and standards, *Online* (©Information Today Inc.), January issue (1999), Available at: http://www.infotoday.com/online/OL1999/milstead1.html (Accessed on 14 August 2008)
13. World Wide Web Consortium (W3C), Available at: http://www.w3.org/Metadata/ (Accessed on 14th August 2008)
14. Hunter J L, A survey of metadata research for organizing the web, *Library Trends*, 52 (2) (2003) 318-344.
15. Smits J, The creation and integration of metadata in spatial data collections, In *Digital Map Librarianship: a working syllabus*, 63rd IFLA Conference, (Copenhagen, Denmark: August 18) (1997)
16. Day M, Metadata: mapping between metadata formats, UKOLN - The UK Office for Library and Information Networking

(University of Bath, UK: May 2002), Available at: http://ukoln.ac.uk/metadata/interoperability/ (Accessed on 27 August 2008)

17. Polydoratou P and Nicholas D, Familiarity with and use of metadata formats and metadata registries amongst those working in diverse professional communities within the information sector, *ASLIB Proceedings*, 53 (8) (2001) 309-324.
18. Op. Cit. Dempsey L and Heery R
19. Quam E, Informing and evaluating a metadata initiative: usability and metadata studies in Minnesota's Foundations Project, *Government Information Quarterly*, 18 (2001) 181-194.
20. DCMI – Dublin Core Metadata Initiative, Available at: http://dublincore.org/ (Accessed on 19 September 2008)
21. DCMI – Dublin Core Metadata Initiative, Dublin Core Metadata Element Set (version 1.1): reference description, Available at: http://dublincore.org/documents/dces/ (Accessed on 19 September 2008)
22. Burnett K, Bor-Ng K and Park S, A comparison of the two traditions of metadata development, *Journal of the American Society for Information Science*, 50 (13) (1999) 1209-1217 [Special issue: Integrating multiple overlapping metadata standards]
23. Dekkers M and Weibel S L, Dublin Core metadata initiative progress report and workplan for 2002, *D-Lib Magazine*, 8 (2) (2002) 1-9, Available at: www.dlib.org/dlib/february02/weibel/02weibel.html (Accessed on 28 September 2008)
24. Guinchard C, Dublin Core use in libraries: a survey, *OCLC Systems and Services*, 18 (1) (2002) 40-50.
25. Johnson B C, XML and MARC: which is right, *Cataloguing and Classification Quarterly*, 32 (2) (2001) 107-126.
26. Library of Congress, MARC standards, Available at: http://www.loc.gov/marc/ (Accessed on 6 October 2008)
27. Library of Congress - MARBI Committee, The MARC 21 formats: background and principles (revised version - November 1996), Available at: http://www.loc.gov/marc/96principl.html (Accessed on 6 October 2008)
28. Kellogg D, Open source OAI metadata harvesting tools (meeting report of Digital Library Federation - October 2004), Available at: http://www.diglib.org/aquifer/oct2504/harvesting.pdf (Accessed on 18 October 2008).
29. Powell A, DC-dot: Dublin Core metadata editor (developed by UKOLN metadata project, University of Bath, UK), Available at: http://www.ukoln.ac.uk/metadata/dcdot/ (Accessed on 2 November 2008).
30. Prasad A R D, Using multiple metadata formats in DSpace, paper presented on 6-8 July 2005 at a User Meet at the Cambridge University, UK.
31. DSpace Federation at MIT, DSpace system documentation: metadata, Available at: http://www.dspace.org/technology/system-docs/functional.html (Accessed on 21 November 2008)
32. Mohamed K, *Merging multiple search results for Meta search engines*, Dissertation submitted at the School of Information Science, University of Pittsburgh, Pittsburgh, 2004.
33. Turner T P and Brackbill L, Rising to the top: evaluating the use of the HTML meta tag to improve retrieval of world wide web documents through Internet search engines, *Library Resources and Technical Services*, 42 (4) (1998) 258-71.
34. Dawson A and Hamilton V, Optimizing metadata to make high-value content more accessible to Google users, *Journal of Documentation*, 62 (3) (2006) 307-327.
35. Mohamed K A F, The impact of metadata in web resource discovering, *Online Information Review* 30 (2) (2006) 155-167.
36. Chan L M, Inter-indexer consistency in subject cataloging, *Information Technology and Libraries*, 8 (4) (1989) 349-358.
37. Weinheimer J L, How to keep the practice of librarianship relevant in the age of the Internet, *Vine*, 29 (3) (1999) 14-37.
38. Thomas C F and Griffin L S, Who will create the Metadata for the Internet?, *First Monday*, 3 (12) (1998), Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/633 (Accessed on 18 December 2008)
39. Greenberg J, Pattuelli M C, Parsia B, and Robertson W D, 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 Applications, (National Institute of Informetrics [NII], Tokyo, Japan: October 24-26) (2001) p.38-46.
40. Richmond A, META tagging for search engines, Web Developer's Virtual Library, Available at: http://www.wdvl.com/Search/Meta/Tag.html (Accessed on 22 December 2008)
41. OCLC – Online Computer Library Center, InterCAT Project, Available at: http://www.oclc.org/research/projects/archive/intercat.htm (Accessed on 28th December 2008)