Metadata Harvesting Tools and Services in Digital Era: A Guide for Professionals

G H S Naidu Prabhat Singh Rajput

Abstract

The paper gives a comprehensive idea about metadata. Describes definition, need and categories. Explains DC and MARC metadata standards and its elements. Discuses the different type of metadata and their functions. Metadata harvester provides indexes or harvests metadata, from different open archives and open access journals. The study attempts to know Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) and available Harvesting services in India.

Keywords: Metadata, Metadata Harvesting, Harvesting Services, Digital library

1. Introduction

The wealth of information and its access provides a frustrating dilemma for librarians and information seekers alike. The information is available, but how to find it, how to organize it, and how to found it again? The users are overwhelmed, with problems when confronted by disorganized and often un-indexed information. Thus the availability of huge sources of unorganized information on the Internet initiated a need to have tools to organize the information, i.e. metadata. Several researches are now engaged in finding ways and means of cataloguing and classifying materials available on the Internet and other online networks. Many metadata schemes have been created by library and information specialists like the MARC format, the AACR-II cataloguing format, subject heading lists such as LC Subject Headings, and Sear's List of Subject Headings and classification schemes such as DDC, UDC and so on. Each of these schemes has been constructed by experts in the relevant field from an understanding of the specific domain, information resources, needs, and the requirements for describing documents.

2. What is Metadata?

Metadata is "data about data". In the context of bibliographic information systems, it is the author, title, place, publisher, subject code, subject heading, etc., for books. In the case of serials, it is the title, publisher, ISSN etc. Similarly, case of a bank account it is name, address, signature, etc. National Information Standards Organization (2004) defines "Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information". [1]

"Metadata is structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities". [2]

The term "metadata" commonly refers to any data that aids in the identification, description and location of networked electronic resources.

The term metadata is used differently in different communities.

- Some use it to refer to machine understandable information, while others use it only for records that describe electronic resources.
- In the library environment, metadata is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital.
- Traditional library catalogue is a metadata tool; MARC 21 and the rule sets used with it, such as AACR-II, are metadata standards.
- Other metadata schemes have been developed to describe various types of textual and non-textual objects, including published books, electronic documents, archival finding aids, art objects, educational and training materials, and scientific datasets.

2.1 Need of Metadata

Metadata is a systematic method for describing resources and thereby improving access to them. The primary aim of metadata is to improve resources discovery.

- Resource documentation
- Resource selection, evaluation and assessment
- Resource identification and location
- Improving the quality and quantity of search result
- Electronic commerce to encode prices, term of pay, etc.
- Protecting instinctual property rights
- Efficient content development and archiving

2.2 Categories of Metadata

Metadata has been divided into five categories as follows:

- **Descriptive metadata:** Include the creator of the resource, its title, subject heading and other elements that will be used to search and locate the items.
- **Structural metadata:** Describe how an item is structured, for examples if it is an electronic book composed of scanned pages, each of which is a separate computer image file.
- **Administrative metadata:** Provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.
- **Rights management metadata:** This is deals with intellectual property rights
- **Preservation metadata**: It is contains information needed to archive and preserve a resource.

2.3 Metadata Standards

Some of the metadata standards available are MARC, MARC21, Dublin Core, UK MARC (now transformed to marc21), etc. MARC21 is the latest standards in term of metadata. The first level metadata elements of MARC are:

- Leader and Directory
- Control Fields 001-008
- Number and Code Fields (01X-04X)
- Classification and Call Number Fields (05X-08X)
- Main Entry Fields (1XX)
- Title and Title-Related Fields (20X-24X)
- Edition, Imprint, etc.Fields (250-270)
- Physical Description, etc. Fields (3XX)
- Series Statement Fields (4XX)
- Note Fields: Part 1 (50X-53X)
- Note Fields: Part 2 (53X-58X)
- Subject Access Fields (6XX)
- Added Entry Fields (70X-75X)
- Linking Entry Fields (76X-78X)
- Series Added Entry Fields (80X-830)
- Holdings, Location, Alternate Graphics, etc. Fields (841-88X)

Among several standards DC is most popular and widely accepted due to its compatibility with almost all kinds of E-sources.

2.4 Dublin Core

DC (Dublin Core) is remarkably different from other metadata standards because of its simplicity, easy to use and interpretability. The DC Metadata Initiatives (DCMI), an international community supported by OCLC, has led to the development of metadata components that enhance cross-disciplinary resource discovery. The mission of DCMI is to develop an easy mechanism for searching and indexing web resources through

- developing metadata standards for cross domain resource discovery;
- defining frameworks for the interpretation of metadata;
- facilitating the development of discipline specific metadata sets that work within the frameworks of cross-domain resource discovery and metadata interpretability.

DC metadata descriptor exists between the crude metadata currently employed by search engines and the complex mass of information encode within records such as those for MARC format. The core element set of DC metadata are as follows:

Title- title of resources	Format- physical or digital
Creator - author	Identifier- URL, ISBN, etc
Subject - subject, keyword	Source- journal article collection, etc.
Description-table of content,	Language- language of resource abstract
Publisher- person/institute	Relation- relationship to other works
Contributor-contributing person/ institute	Coverage- geographic/temporal coverage
Date- date	Right- copyright date, etc.
Type- nature of content	

2.5 Types of Metadata and their Functions

Type	Definition	Examples
Administrative	Metadata used in managing and administering information resources	Acquisition information Rights and reproduction tracking Documentation of legal access requirements Location information Selection criteria for digitization Version control and differentiation between similar information objects Audit trails created by record keeping systems
Descriptive	Metadata used to describe or identify information resources	Cataloguing records Finding aids Specialized indexes Hyperlinked relationships between resources Annotations by users Metadata for record keeping systems generated by records creators
Preservation	Metadata related to the preservation management of information resources	Documentation of physical condition of resources Documentation of actions taken to preserve physical and digital versions of resources, e.g., data refreshing and migration
Technical	Metadata related to how a system functions or metadata behave	Hardware and software documentation Digitization information, e.g., formats, compression ratios, scaling routines Tracking of system response times Authentication and security data, e.g., encryption keys, passwords
Use	Metadata related to the level and type of use of information resources	Exhibit records Use and user tracking Content re-use and multi-versioning information

3. Open Access Initiative for Metadata Protocol Harvesting

Open Access works are scattered across many disciplinary archives, institutional e-print archives, institutional repositories and open access journals. Therefore, it is difficult for users to locate all

needed works on a particular subject. One important international movement to solve this problem is the Open Archives Initiatives (OAI), which aims to develop and promote the use of a standard protocol, know as the Open Archives Metadata Harvesting Protocol (OAMHP), designed for better sharing and retrieval of e-prints residing in distributed archives.

4. Metadata Harvesting Services in India

(i).	**Name:**	Search Digital Libraries (SDL)
	URL:	http://drtc.isibang.ac.in/sdl
	Host:	DRTC Bangalore
	Software Used:	PKP (Public Knowledge Project)

Description: The SDL currently has **20130** papers from **9** archive(s) indexed and compatible with versions 1.1 and 2.0 of the OAI Harvesting Protocol. The PKP Open Archives Harvester is a free metadata indexing system and federally funded efforts to expand and improve access to research. The PKP OAI Harvester allows you to create a searchable index of the metadata from Open Archives Initiative-compliant archives, such as sites using Open Journal Systems or Open Conference Systems. It indexes Australian Library and Information Science Association (ALIA); CNR Bologna Research Library, Italy; Dialogo Cientifico utilize, Brazil; DLIST, University Arizona; DSPACE inra Avignon; E-LIS: E-Prints in Library and Information Science; Subject Gateway of Library and Information Services etc.

(ii).	**Name:**	Knowledge Harvester@INSA
	URL:	http://61.16.154.195/harvester/
	Host:	INSA
	Software Used:	PKP (Public Knowledge Project)

Description: Knowledge Harvester@INSA is an experimental initiative from INSA (Indian National Science Academy), which currently has 2,011 papers from 3 archives indexed. It indexes African Journals Online, European Integration, INSA Digital Library.

(iii).	**Name:**	Open J-Gate
	URL:	www.openj-gate.com
	Host:	Informatics (India) Ltd.

Description: Open J-Gate is an electronic gateway to global journal literature in open access domain launched in 2006. Open J-Gate is the contribution of Informatics (India) Ltd to promote OAI. Open J-Gate provides seamless access to millions of journal articles available online. Open J-Gate is also a database of journal literature, indexed from 4373 open access journals. Out of them 1,500+ are peer reviewed scholarly journals.

(iv). **Name:** SJPI Cross Journal Search Service
URL: http://144.16.72.144/harvester/
Host: NCSI, IISc
Software Used: PKP (Public Knowledge Project)

Description: The SJPI Harvester currently has 1,047 papers from 13 journals indexed. It indexed Bulletin of Materials Science; Currently Science; Journal of Astrophysics and Astronomy; Journal of Biosciences; Journal of Chemical Science; Journal of Genetics; Journal of the Institute of Science; SRELS Journal of Information Management etc.

(v). **Name:** SEED (Search Engine for Engineering Digital-Repositories)
URL: http://eprint.iitd.ac.in/seed/
Host: IIT, Delhi
Software Used: PKP (Public Knowledge Project)

Description: The Seed currently has 6,176 papers from 4 archives indexed. It indexed Dspace@NITR; Earthquake Engineering; Eprint@IISc; Eprint@IIT Delhi.

5. Conclusion

Metadata is an essential tool, which be developed as a standard in this digital era as guide for libraries and library professionals. Among many promises of global information infrastructure is to improve access to information, regardless of the location, format and structure of the resources. The concept of "The Semantic Web" is possible if researchers from different fields will come together and agree upon one metadata standard schema. Then new areas of interest and research can build on it, improve and expand it. It will take long time effort to achieve this kind of uniformity. There is a need for cross-walks between various emerging metadata sets to avoid enormous problems with interoperability in the future content provide will be interested in making appropriate use of metadata.

References

1. Niso. Understanding Metadata. Available at http://www.niso.org/standards/resources/UnderstandingMetadata.pdf (Accessed on 26.12.2007)
2. Hode, Gall. Metadata Made Simple: A guide for Libraries. Available at http://www.niso.org/news/Metadata_simpler.pdf (Accessed on 05.01.2008)
3. MarcC21 Concise Format for Bibliographic Data. Available at http://www.loc.gov/marc/bibliographic/ecbdhome.html (Accessed on 12.01.2008)
4. Taskforce on Metadata. Committee on Cataloguing: Description and Access. Available at http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html. (Accessed on 05.01.2008)
5. Search Digital Libraries. Available at http://drtc.isibang.ac.in/sdl/ (Accessed on 05.01.2008)
6. Gins, Jean. AGNEW, Grace. and ELIZABETH, Brown. Getting mileage out of metadata: Applications for the library. Chicago: American Library Association, 1999. pp.1.

7. Lazinger, Susan S. Digital Preservation and Metadata: History, theory, practice. Englewood: Libraries Unlimited, 2001. pp. 139-188.

8. Hirwade, Mangala and HIRWADE, Anil. Metadata Harvesting Services in India. Library Herald. 2004, (4), pp. 275-281.

9. Chudamani, K. S. Metadata and Content Management in the Digital Environment. SRELS Journal of Information Management. 2005, (2), Pp. 207-211.

10. Kapoor, Kanta. Metadata: A pathway to electronic resources. Annals of Library and Information Studies. 2002, 49 (1), pp. 7-11.

About Authors

Dr. G H S Naidu, Librarian & Head, School of Library & Information Science, Devi Ahilya Vishwavidyalaya, Indore (M.P.).
E-mail : head.cl@dauniv.ac.in.

Mr. Prabhat Singh Rajput, Lecturer, School of Library & Information Science, Devi Ahilya Vishwavidyalaya, Indore (M.P.)
E-mail : prabhat.t82@gmail.com.