Indian Journal of Library and Information Science
Volume 8 Number 2, May - August 2014

Original Article

Metadata Harvesting: Digital Preservation towards Open Access Initiatives (OAI) in India

C. Baskaran*, M. Sadik Batch**

Abstract

This papers describes that metadata harvesting for Digital preservation its provides indexes or harvests metadata from different archives and digital documents to be made as institutional repositories .The functions and creations of metadata into different facets and the Dublin Core metadata initiatives an information community supported by OCLC has led to the development of metadata components that enhance cross disciplinary resource discovery. It is also provides some of the international level meta data harvester and web portal on knowledge Harvester @NSA open gate, SJPI cross journal search service, SEED (Search engine for Engineering Digital repositories) PKP (Public Knowledge Project).

Keywords: Metadata; Digital preservation; Dublin core; OAI-PMH; PKP; SEED.

Introduction

The term metadata 'data about data', i.e. a set of information which remains in same intentional hierarchical relationship with another set of information. So it is a summary of data about some other data. Another concept is that "it is machine understandable information for the web". Presently the term refers to any data used to aid the identification, description and location of networked information. For example, a metadata system common in libraries the library catalogue – contains a set of metadata records with elements that describe a book or other library item: author, title, data of creation or publication, subject coverage, and the call number specifying location of the item on the shelf. Metadata enhances retrieval performance – Metadata can improve retrieval by establishing a context for individual descriptors. For example the word 'Green' in the Creator or Author field indicates the name of an individual, where as 'Green' in the title of a document may be a subject retrieval term.

Author's Affilation: *Deputy Librarian, School of Library and Information Science, Alagappa University, Karaikudi - 630003, Tamil Nadu, India, **Associate Professor, LIS, wing, DDE, Annamalai University, Annamalainagar - 608002, Tamil Nadu, India.

Reprint's Request: C. Baskaran, Deputy Librarian, School of Library and Information Science, Alagappa University, Karaikudi - 630003, Tamil Nadu, India.

E-mail: cbklis@gmail.com

Digital Preservation

The preservation of digital information can be seen as one of the greatest challenges for the library and information professions at the end of the twentieth-century. It is easy to get caught-up in the general enthusiasm for digital libraries but more consideration needs to be given to the problem of making this information available to future generations. Useful work has been published by the US Commission on Preservation and Access (CPA), including an important report by a Task Force on the Archiving of Digital Information (TFADI) jointly commissioned with the Research Libraries Group.

Libraries have traditionally understood at least one of their roles as the preservation of information for future use. This has been especially true of national libraries and selected institutions in the research library sector. For example, in early 1996 the British Library proposed that legal deposit should be

extended to include non-print materials, even allowing for the possibility of collecting networked (on-line) publications if this became technically and economically feasible. Most countries in Europe, North America and Australasia have either extended legal deposit to electronic publications or are considering doing so. This implies a commitment, by national libraries at least, to the long-term preservation of digital publications.

The main problem with digital preservation is that digital technology, in comparison to print, is an extremely fragile medium for the cultural memory of the world. The most commonly given example of this fragility is the 1960 United States Census, where raw data stored on magnetic tapes apparently became obsolete and, to all intents and purposes, unreadable by the late nineteen-seventies.

Digital Information: Demerits,

1. The storage medium - digital storage media, whether magnetic or optical, are subject to relatively rapid decay: especially when compared with print.
2. The hardware and software - digital information is machine-dependent, and to be 'read' accurately it needs specific computer hardware and software. Unfortunately, hardware and software quickly become obsolescent or otherwise unusable.

Proposed solutions to these problems usually involve periodic 'refreshing' or recopying of the digital information onto new media and the occasional 'migration' of data into new formats. Assuming that some answer can be found to these problems, there remains the important issue of intellectual preservation. Even when digital information has migrated into new formats, there will remain a need for users to be sure that the 'document' they are looking at is the one that they were looking for.

The archives community - especially in the United States - has been addressing these issues for some time now. This has partly been led by the need for electronic records to be accepted in legal evidence but the fact that electronic records are increasingly created in a variety of important situations: government; health-care; business-transactions, etc., has resulted in a renewed interest in authentication and validation issues. A research project at the University of Pittsburgh School of Library and Information Science has been investigating "Functional Requirements for Recordkeeping" and has attempted to identify and specify the fundamental properties of records. The functional requirements identified by the project emphasized that records should be comprehensive, identifiable, complete, authorized, preserved, removable, exportable, and accessible and redact able. The project suggested that records (or 'Business Acceptable Communications') should carry a six layer structure of metadata which would contain not only a 'Handle Layer' (including a unique identifier and resource discovery metadata) but also very detailed information on terms & conditions of use, data structure, provenance, content and the use of the record after its creation. The metadata is intended to carry all the necessary information that would allow the record to be used - even when the "individuals, computer systems and even information standards under which it was created have ceased to be" (Ibid).

Definition about Meta data

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, and 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. 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.

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". Metadata is a systematic method for describing resources and thereby improving access to them.

Creation of Metadata

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

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-resources.

Dublin Core and its Implication

DC (Dublin Core) is remarkably different from other metadata standards because of its simplicity, easy to use and interoperability. 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.

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.

Metadata Harvesting Services in India

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.

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.

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.

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.

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

D-space @NITR; Earthquake Engineering; E-print @IISc; E-print @IIT Delhi.

Conclusion

The Meta data harvesting is the significant process into digital preservation, which highly involves in Institutional repositories in India as well as global perspectives. Transferring information on web technology using basic need of academic society, which is predominant factor requirement for preservation of digital documents and scholarly information need to the Higher Education. The simplicity semantic interoperability (cross domain), international consensus, extensibility and modularity have become the qualities of DC. Cultural heritage and information Professional such as museum registrars, library cataloguers, and archival processor are increasingly applying the term metadata to the value-added information that they create to arrange, describe, track and otherwise enhance access to information objects, carefully designed metadata results in the best information management in the short and long term. It is now a viable option and hence found widespread acceptance among the electronic information community.

References

1. Rothenberg, J. Ensuring the longevity of digital documents. *Scientific American.* 1995; 272(1): 24-29.
2. Day, MW. Preservation of electronic information: a bibliography. 1997. http://www.ukoln.ac.uk/~lismd/preservation.html
3. Task Force on the Archiving of Digital Information. Preserving digital information: report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group. Washington, D.C.: Commission on Preservation and Access; 1996. http://www.rlg.org/ArchTF/
4. British Library Research and Innovation Centre. Proposal for the legal deposit of non-print publications. London: British Library; 1996. http://portico.bl.uk/ric/legal/legalpro.html
5. Weinberg, GL. The end of Ranke's history? Reflections on the fate of history in the twentieth century. In: Weinberg, GL. Germany, Hitler, and World War II: essays in modern German and World history. Cambridge: Cambridge University Press; 1995, 325-336.
6. Mallinson, JC. On the preservation of human- and machine-readable records. *Information Technology and Libraries.* 1988; 7: 19-23.
7. Graham PS. Intellectual preservation: electronic preservation of the third kind. Washington D.C.: Commission on Preservation and Access. 1994. http://www-cpa.stanford.edu/cpa/reports/graham/intpres.html
8. Piasecki SJ. Legal admissibility of electronic records as evidence and implications for records management. *American Archivist.* 1995; 58(1): 54-64.
9. Bearman, D and Sochats, K. Metadata requirements for evidence. Pittsburgh, Penn.: Archives and Museum Informatics; 1996.

http://www.lis.pitt.edu/~nhprc/BACartic.html

10. Rothenberg, J. Metadata to support data quality and longevity. 1996. http://www.computer.org/conferen/meta96/rothenberg_paper/ieee.data-quality.html
11. Naidu, GHS and Prabhat Singh Rajpat. Metadata Harvesting tools and Services in Digital Era: A Guide for Professionals. *6th International CALIBER-2008, University of Allahabad.* Feb-28-1 March,2008.
12. NISO. Understanding Metadata available at http://www.niso.org/standards/resources
13. Dublin core Metadata initiatives http://www.dublincore.org