Developing Virtual Union Catalogue of ETDs on Health and Medicine: A Practical Approach

Prosenjit Sarkar[1] and **Dr. Parthasarathi Mukhopadhyay** [2]

[1]**Superintendent (Library Services)**
Central Library, The University of Burdwan, Burdwan-713104
Email: prosenjit.toton@gmail.com

[2]**Assistant Professor**
Department of Library and Information Science
The University of Burdwan, Burdwan – 713 104
Email: psmukhopadhyay@gmail.com

Keywards: ETD, ETD union catalogue, Interoperability, Metadata harvesting, OAI-PMH.

Abstract: This paper reports designing of a prototype union catalogue of ETDs on health and medicine through the application of metadata harvesting from OAI-PMH compliant ETD repositories and in turn this union catalogue will act as a single-window search interface to facilitate retrieval of ETDs on global scale and to encourage resource sharing in the digital arena.

Introduction

ETDs stand for Electronic Theses and Dissertations. Generally digitally/electronically submitted Theses and Dissertations (TDs) are called ETDs. TDs are the intellectual outputs of scholars' and represent academic heritage of an institution. TDs in paper format and Electronic Theses and

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

Dissertations (ETDs) in electronic/digital formats are valuable academic resources. Therefore, it is essential to archive, preserve and disseminate these intellectual outputs. But these valuable resources are underutilized due to lack of suitable access system. Since long time, libraries have stored and circulated these TDs through manual processes and in most of the libraries these resources are lying obscure. Efforts have been increasing across the globe to enhance accessibility and visibility of TDs. Library automation and Internet has changed the whole scenario of academic sector including the management of TDs and as a result higher academic institutions are developing their own ETD systems. At present, a large number of institutions are investing money, man-hours and efforts in creating, archiving and disseminating ETDs in institutional repositories. These ETD systems are developed by using different software (e.g. DSpace, EPrint, Fedora, Greenstone etc.). But without a good retrieval platform, full texts and even citations of these scholarly works will not be accessible through different digital channels including Internet [1]. Universities including medical institutes are in the process of opening up their treasures of knowledge which are in the form of ETDs. But without access to recent international scholarly works and without the ability to contribute to national research results, scholars of Medical sciences in the developing

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

countries are struggling to make an impact in their discipline [2]. The development of interoperable ETD systems in Medical sciences may solve this problem; these ETD systems are designed and developed by using different software platforms, different metadata schemas/content designators and different retrieval models. Fortunately Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) compliant open source software have solved this problem. Moreover installation bases of these software are commendable. These archiving software differ significantly in scope and retrieval features. So, users need to be accustomed with retrieval features of these digital archiving systems used in building ETD repositories in health and medicine. In this context, authors of this research report have designed a prototype union catalogue of ETDs on health and medicine as a single-window search service to facilitate retrieval of ETDs on health and medicine on global scale. This prototype is named as UniMed and service is available through high-speed network connectivity in the Intranet environment of the University of Burdwan. UniMed is designed by automatic metadata harvesting from distributed ETD systems, through the application of OAI-PMH version 2.0 [3]. The aim of this union catalogue is to build an easy-to-use search service for providing access to databases of worldwide ETDs along with

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

scholarly literature on health and medicine and foster resource sharing in the digital library system.

ETD

ETD is an acronym of Electronic Thesis and Dissertation. It is a new genre of document. ETD is digital document which is similar to its paper printed counterpart, only difference is, ETD is prepared digitally. Generally, there are two types of ETD – first one is digitally/electronically prepared and submitted TD and second one is scanned paper printed TD, which is called retrospective conversion. According to UNESCO ETD Guide, "An ETD is a document that explains the research or scholarship of a researcher/student. It is expressed in a form simultaneously suitable for machine archives and worldwide retrieval. The ETD is similar to its paper predecessor. It has figures, tables, footnotes and references. It has a title page with the author's name, the official name of the University, the degree sought, and the names of the committee members. It documents the author's years of academic commitment. It describes why the work was done, how the research relates to previous work as recorded in the literature, the research methods used, the results, and the interpretation and discussion of the results, and a summary with conclusion." (http://etdguide.org/).

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

The advantages of ETDs over paper printed TDs are numerous and also discussed by several authors such as [4] [5] [6]. These are stated below:-

- ETD submission has less cost in comparison to submit paper printed TDs;
- ETD prevent duplicate efforts in research works;
- ETDs are more frequently and simultaneously accessed by multiple users at a time at 24×7 modes but increased use does not require extra library staff;
- Highly expressive multimedia application is possible in ETDs;
- ETDs helps academic institutions to implement digital library services;
- Implementing ETD systems enhance status and prestige of the Library; and
- ETDs save library space as well as money and man-hours by eliminating much manual process.

Objectives of the study

The objectives of this study are:-

- to design a prototype union catalogue of ETDs on health and medicine by using open source software and open standards;

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

- to facilitate retrieval of ETDs on health and medicine on global scale by single-window search service;
- to encourage designing of OAI-PMH-compliant interoperable ETD systems;
- to encourage resource sharing in the digital library environment; and
- to promote access to ETDs through enhancing visibility.

Scope

OpenDOAR (Directory of Open Access Repositories) (see http://www.opendoar.org/find.php?search=&cIID=&ctID=6&rt&ID=&cID=&IID=25&rSoftWareName=&submit=Search&format=summary&step=20&sort=r.rName&rID=&ctrl=new&p=1) was consulted for selecting ETD IDRs on health and medicine. Selections of ETD IDRs are based on carefully crafted criteria, that is IDRs must have English language ETDs and should have valid OAI-PMH base url. An exhaustive review and harvesting of existing ETD repositories on health and medicine is out of the scope of this paper. It is a prototype archive.

OAI

OAI is an acronym of the Open Archives Initiative. The Open Archives Initiative (OAI) is an international consortium, formed by a broad range of

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

researchers, librarians, publishers, and archivists whose ultimate aim is to create simple standards to support interoperability among digital repositories by metadata harvesting [7]. The OAI is dedicated to solving problems of digital library interoperability [8]. 'Open Archive' defines information access at free of cost and without any restriction (with proper acknowledgement). The Open Archive Initiative means an Open Archive, which implements OAI Protocol for Metadata Harvesting to allow remote archive to access its metadata using an open standard. A Repository or IDR is synonymous to an Open Archive. In the context of OAI, a Repository or IDR should be network accessible that supports OAI-PMH.

The roots of OAI work lies in an effort to enhance access to e-print repositories. OAI can be mentioned as a lightweight transport mechanism to expose and disseminate metadata, based on standard technology for transmission of contents and discovery [9].

Interoperability

Interoperability refers different aspects of archive initiatives e. g. metadata formats, architectural framework, and their openness to create third-party digital library services, their integration with current process of scholarly

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

communication, usability in cross-disciplinary context etc. [10]. Interoperability among repositories provides benefits to the scholars that use them.

Some definitions of Interoperability are given below: -

"Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality" [11].

"Interoperability is the ability of two or more systems or components to exchange information and use the exchanged information without special effort on either system" [12].

In general, 'Interoperability' can be defined as the compatibility of two or more different computer systems so that they can exchange data and information and can use the exchanged data and information without any kind of manipulation or loss. There are several methods to achieve interoperability; building union catalogue by implementing OAI-PMH is one of them.

OAI-PMH

In the context of digital library, implementation of OAI-PMH enhance discovery of Web-based resources and extend the aim of traditional descriptive cataloguing in the digital environment. The OAI-PMH released in an attempt to

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

address interoperability problem among the many existing and independent IDRs. The aim of this Initiative was on high-level communications among repositories and the simplicity of protocols [13]. OAI-PMH has two components - Data Provider, Service Provider. It also defines six 'verbs' (e. g. Identify, ListMetadataFormats, ListSets, GetRecord, ListIdentifiers and ListRecords); these 'verbs' are used as syntaxes to interact between these two components. Data Provider exposes its metadata for harvesting and Service Provider harvests exposed metadata and stores it locally and on the basis of the harvested metadata provide services to end users. A Data Provider archive must satisfy some basic requirements - the archive must have an online interface and a web server that can be used for the purpose of protocol. The basic purpose of this protocol is incremental bulk transfer of metadata which is called harvesting and harvested metadata are used as resource discovery tools. Here, the system attempts to design union catalogue (a single-window search service) of ETDs on Health & Medicine through harvesting metadata from participating ETD IDRs. It is important to mention here that this union catalogue only provides access to metadata – it does not provide the resources themselves. However, every retrieved record against a search displays hyperlink to full-text object (available in the archive where it uploaded) along with bibliographical information. The union catalogue archive of

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

ETDs on health & medicine is designed to function as both Service Provider and Data Provider. The interaction between Service Provider and Data Provider may be represented schematically as in Fig.1.

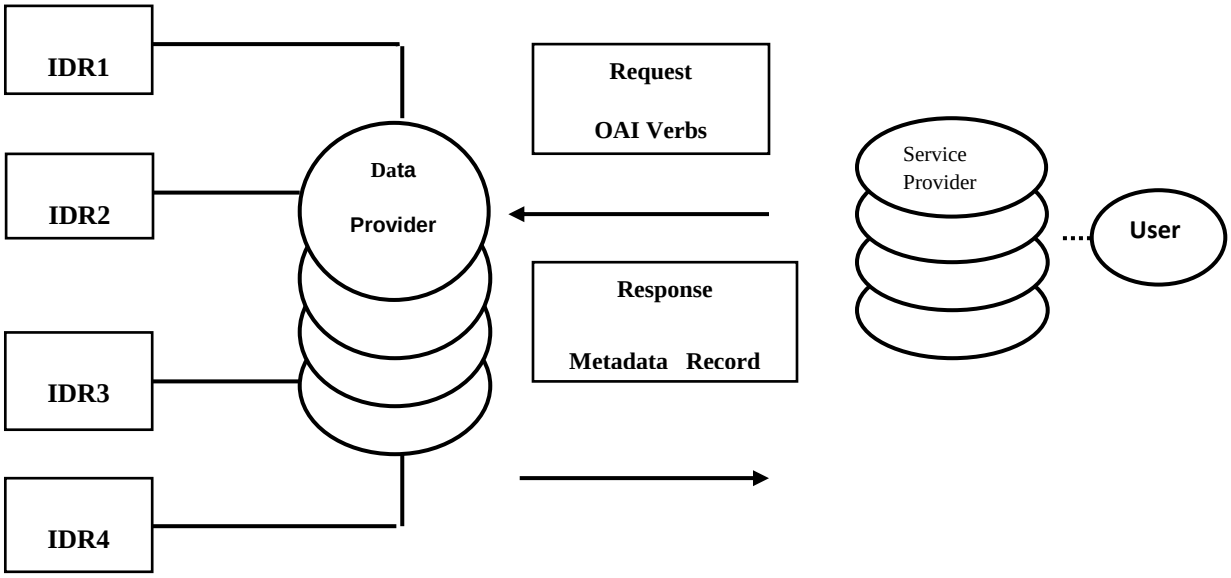

Fig. 1- Functions of Data Provider and Service Provider in OAI-PMH Framework

Designing harvesting framework for union catalogue

The following steps are required for design and development of harvesting framework for health and medicine ETDs. There are three major steps -

i) Software framework development; ii) Selection, installation and configuration of harvesting tool; and iii) Selection of ETD repositories and

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

collection of essential attributes for harvesting (name of the IDR, resource URL, base URL, etc.); PKP has been used as harvester to design this system.

The prototype harvesting framework for ETDs on health & medicine, named as UniMed, is based on open source software and open standards. This framework is based on Linux-Apache-MySQL-PHP/PERL (LAMP) architecture. Details of designing harvesting framework already reported by the authors in their article published in journal *Annals of Library and Information Studies* in the December issue of 2010 [14].

As a whole, the use of the open source software in developing the framework depends on a structured methodology. The steps related to the process of the designing harvesting framework may be divided into three major groups [15]. These are as follows –

Group A: LAMP related activities;

Group B: Harvester related functions.

This group includes two major steps:

- Installation of PKP harvester.
- Configuration of PKP harvester.

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

Group C: Repository related activities.

The most important task of the administrator is to set up archive(s) for metadata harvesting. The UniMed started with fourteen OAI-PMH compatible open access IDRs containing ETDs of health and medicine. ETD repositories were selected on the basis of carefully crafted criteria; i. e. IDRs should have ETDs in English language; and must have valid OAI-PMH base URL. For this research project, OpenDOAR was consulted for selecting ETD IDRs, and collected following information e. g. name of the repository, their URL and OAI-PMH base URL. At present, OpenDOAR listed seventy ETD repositories on health and medicine; out of seventy, forty six repositories are of English language and thirty two ETD repositories (English language) have OAI-PMH base URL. Out of forty six English languages ETD repositories continent wise position of ETD systems on health and medicines are as follows – Australasia four; Asia eleven; Europe twenty one; North America ten. In Asia, total ETD repositories are fourteen and out of fourteen repositories eleven possess English language ETDs. Asian country-wise distribution of English language ETD systems are as follows – Korea Republic one; Bangladesh one; Taiwan two; Japan three; China one; Indonesia one and India two.

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

UniMed has so far harvested metadata from fourteen IDRs and agglomerated a total of more than seventy three thousand (73,000) open access scholarly materials including ETDs of health and medicine (due to space constraint details of harvested ETD repositories are not listed here).

Some snapshot of UniMed is given below. Fig. 2 is a snapshot of metadata harvesting. Fig. 3 is a snapshot of search queries. It supports localized searching in two modes – simple and advance. End-user can limit in a single repository or group of repositories. Search can be filtered by Dublin Core Metadata Element Sets like title, author date range etc. Fig. 4 is a snapshot of browsing ETD archives of health & medicine. The harvesting was conducted during the period from 16th August to 15th September 2011.

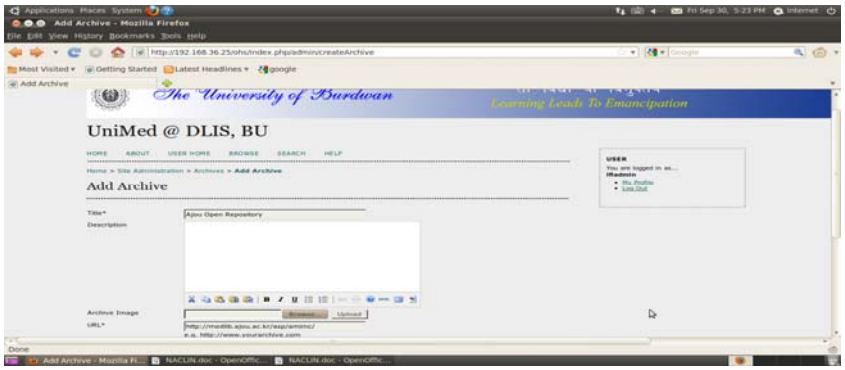

Fig. 2 - A snapshot of harvesting

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

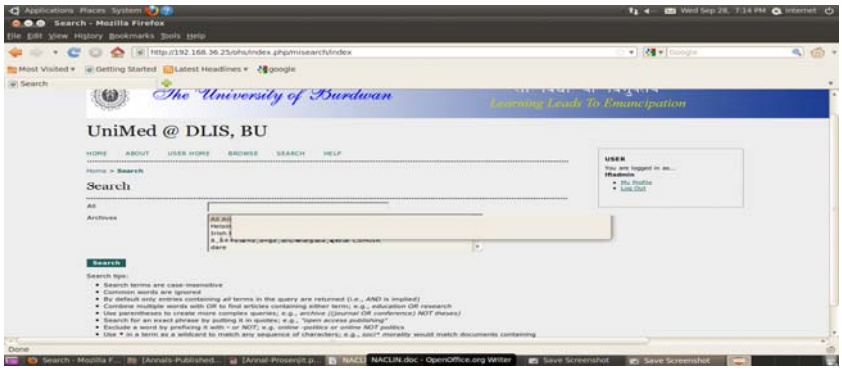

Fig. 3 - A snapshot of search queries

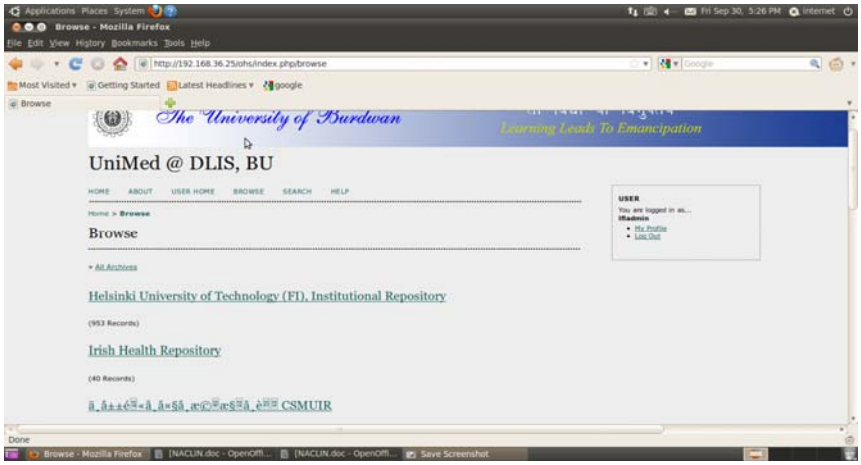

Fig. 4 - A snapshot of browsing health and medicine archives

Conclusion

UniMed is a localized single-window resource discovery tool for ETDs that can harvest and update metadata from various OAI-PMH compliant open access ETD systems on health and medicine. This system has the capability

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

to include new OAI-PMH compitable ETDs of any kind of subject and update existing records for a comprehensive single platform local search service open access IDRs containing ETDs on health and medicine and foster resource sharing in the digital environment. Limitations of this study are that, in spite of having OAI-PMH base URL, harvesting of some repositories was not possible though repeated efforts were made. This may be because of invalid OAI-PMH base URL or may be the non-availability of required data transfer speed in network domain.

References

[1] Wang, I. S, and Y -O Li. (2009). Creating a virtual union catalog for Hong Kong dissertations and theses collection. The Electronic Library, 27(2), 331-341.

[2] Kirsop, B. (2008). Open access to publicly funded research information: the race is on. DESIDOC Journal of Library and Information Technology, 28(1), 41-48.

[3] Lagoze, C, Herbert Van de Sompel, M. Nelson, and S. Warner. (Eds). (2002). The Open Archives Initiatives Protocol for Metadata Harvesting: Protocol Version 2.0 of 2002-06-14. Retrieved October 14, 2009, from http://www.openarchives.org/OAI/openarchivesprotocol.html

[4] Fineman, Y. (2003). Electronic Theses and Dissertations. portal: Libraries and the Academy, 3(2) , 219-227.

[5] Fox, E. A., McMillan, G., Suleman, H., Goncalves, M. A. & Luo, M. (2004). Networked digital library of theses and dissertations (NDLTD). In J. Andrews, & D. Law, (Eds), Digital libraries : policy, planning and practice (pp. 167-187) Aldershot: Ashgate.

[6] McMillan, G. (2004). Implementing ETD services in the library. In E. A. Fox, S. Feizabadi, J. M, Moxley, C. R. Weisser, (Eds), Electronic theses and dissertations : a sourcebook for educators, students and librarians (pp. 319-329). New York: Marcel Dekker.

[7] Suleman, H. (2002).Open digital libraries. Retrieved July 9, 2009, from http://scholar.lib.vt.edu/theses/available/etd-11222002-155624/unrestricted/odl.pdf

[8] Suleman, H. and E. A. Fox. (2002). The Open Archives Initiatives. Journal of Library Administration 35(1-2), 123-145.

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0

[9] Mischo, W. H. (2004). United States federal support for digital library research and its implications for digital library development. In J. Andrews and D. Law (Eds.), Digital libraries : policy, planning and practice (pp. 5-17). Aldershot: Ashgate.

[10] Sompel, Herbert Van de. and Lagoze, C. (2000). The Sante Fe Convention of the Open Archive Initiative. Retrieved October 22, 2009, from http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

[11] NISO (National Information Standards Organization). (2004). Understanding metadata. Retrieved October 14, 2009, from http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

[12] CCDA (Committee on Cataloging: Description & Access). (2000). Association for Library Collections & Technical Services, Task Force on Metadata : Final report. June 16, 2000. Retrieved October 14, 2009, from http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html

[13] Sompel and Lagoze Op. Cit.

[14] Sarkar, P. and Mukhopadhyay, P. (2010). Designing single-window search service for electronic theses and dissertations through matadata harvesting. Annals of Library and Information Studies, 57(4): 356-364.

[15] Mukhopadhyay, P. (2006). Five laws and ten Commandments: the open road of library automation in India. Retrieved May 10, 2009, from http://drtc.isibang.ac.in/handle/1849/409

[16] Mukhopadhyay, P. (2004). Measuring Web Impact Factors: A webometric study based on the analysis of hyperlinks. *Proceedings of the IASLIC XXI National Seminar on Information Support for Rural Development* (pp. 411–425).

[17] Mukhopadhyay, P. (2005). Use of FRBR as model of bibliographic description in online environment. *Vidyasagar University Journal of Library and Information Science*, 51–69.

Preprint of the following paper:

Sarkar, P. and Mukhopadhyay, P. (2012). Developing virtual union catalogue of ETDs on health and medicine : a practical approach. *Proceedings of the National Seminar on Challenges in Library Management System (CLMS 2012).* Indian Association for the Cultivation of Science, Kolkata. pp 334-340. ISBN : 978-81-923191-0-0