Designing Single-Window Search Service for Electronic Theses and Dissertations through Harvesting

Prosenjit Sarkar, Superintendent (Library Services)  
Central Library, The University of Burdwan, Burdwan-713104,  
prosenjit.toton@yahoo.co.in, prosenjit.toton@gmail.com

Dr. Parthasarathi Mukhopadhyay, Sr. Lecturer  
Department of Library and Information Science  
The University of Burdwan, Burdwan – 713 104  
parthasarathimukhopadhyay@gmail.com

INTRODUCTION:

Teaching and research are two important aspects of a university education system. Therefore, it is essential to archive, preserve and disseminate the intellectual output of a university. In this context, theses and dissertations may be viewed as important showcases of a university’s achievement and contribution in higher education. At present, a large number of universities and higher learning institutes are investing money and efforts in creating, archiving and disseminating the citations as well as theses and dissertations including ETDs (Electronic Theses and Dissertations) in institutional repositories. Obviously without a good search platform, full texts and even citations of these researchers’ scholarly works will not be accessible through different digital channel including Internet. The web-based institutional repositories are distributed across the globe and using different software solutions. Fortunately, there is an array of OAI-PMH compatible open source software for archiving digital knowledge objects (e.g. DSpace, E-Print archive, Fedora, Greenstone etc.). These digital media archiving software differ considerably in scope and retrieval functionalities. As a result, users suffer a lot in terms of the load of learning retrieval features of different software. In this regard, this system has designed a prototype union catalogue of electronic theses and dissertations (ETDs), as a single-window search interface to facilitate the retrieval of electronic theses and dissertations on global scale. The

Preprint version of the following paper

aim of this union catalogue is to build an easy search service model using automatic record harvesting to provide access to database of worldwide institutional repositories of ETDs. Of course the repositories need to be compatible with interoperability standard(s) for building a search service on the basis of harvested metadata. Thus we considered interoperability is the process of resource discovery and access. The OAI’s Protocol for Metadata Harvesting is a robust solution of interoperability problem, which defines a standard method of metadata exchange.

**UNION CATALOGUE OF ETDs:**

A union catalogue is a combined library catalogue describing the collection of a number of libraries (see http://en.wikipedia.org/wiki/union_catalog). Union catalogues have been created in a variety of media – networked electronic database is one of them. Union catalogues are useful to librarians as well as researchers, because they help in locating and requesting resources from other libraries, which is a part of library cooperation. Union catalogue of ETDs may assist researchers or library users to search collections of distributed ETDs across the globe. The visibility of union catalogue on ETD may improve considerably with its online availability through the Web. The online union catalogue of ETDs is a system that discovers, organizes and stores metadata records from heterogeneous IDRs, and thereby provides a single-window access interface for distributed ETDs.

There are two processes to build a union catalogue, i.e., virtual (distributed) and centralized. In this project, centralized approach has been used to build ETD union catalogue, which is based on Open Achieve Initiative (OAI) framework. The IDRs containing ETDs are different in their subject coverage, software usage, nature of contents and most importantly in retrieval techniques and tools. As a result it is difficult for end users to search these repositories comprehensively that provide free access to ETDs. This situation demands for the development of

Preprint version of the following paper

a single window search service covering all the IDRs. This single window search service (based on metadata harvesting) is advantageous to researchers and others as it brings them closer to uniform access interface for scholarly information bearing objects².

**BACKGROUD OF OAI CONCEPTS:**

At present time, the Internet has made easy and rapid dissemination of information. Internet has given ample opportunities to research scholar to share their research result among the colleagues in their subject field. But the traditional publication process (in print medium) slow downs the information dissemination or exchange. The Internet has solved this problem. The revolution in Information and Communication Technologies and the need for rapid dissemination of scholarly literature have forced Higher Academic and Research Institutions to build e-print archive or digital library. On the other hand World Wide Web created information management problem, but there was no simple solution to solve this problem. The problem includes persistence, reliability and authenticity¹⁴. The most important issue was the perpetual access. Another problem is the issue of authenticity of a source producing digital objects. These and some other problems forced Higher Academic and Research Institutions to build e-print archives or repositories of digital objects that called Institutional Digital Repository (IDR) with some enhanced service to the user and as a result thousands of IDRs emerge all over the World. Most of these IDRs developed arbitrarily. The concept of Open Archive Initiative was lying with the emergence of increasing number of e-print archives.

The concept “Open Archive” defines information access at free of cost and without any restriction. The Open Archive Initiative means an Open Archive, which implements OAI Protocol

Preprint version of the following paper

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 OAI is an international effort to enhance access to e-print archives and also a current important topic in digital library research and implementation. OAI can be regarded as a lightweight transport mechanism to expose and disseminate metadata, as a standard-based technology for document transmission and discovery, and changed component in the changing scholarly communication framework.

The network connected IDRs have been built arbitrarily in isolation using different software platform, technologies, protocol, and data structures. These differences prevent in the development of digital library services that enable users to use single search interface to discover resources from multiple Institutional Digital Repositories. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is an international effort to address this interoperability problem among distributed IDRs.

**INTEROPERABILITY AND CROSSWALK:**

Interoperability refers to many diverse aspects of archive initiatives. It includes – metadata formats, architectural framework, their openness to create third-party digital library services, their integration with existing process of scholarly communication, usability in cross-disciplinary context etc. In the technical arena it is supported by open standards for communication between systems and for description of resources and collections, among others.

There are many definitions of Interoperability have been given. Some examples are given below:

---

Preprint version of the following paper

“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”8.

“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”8.

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. Crosswalks facilitate interoperability. A crosswalk is "a mapping of the elements, semantics, and syntax from one metadata scheme to those of another”8. At present, crosswalks are common method to achieve interoperability between or among metadata schema. It allows metadata created by one community to be used by another group that employs a different metadata standard. Interoperability and crosswalk ensures exchange of bibliographic data among heterogeneous IDRs across the globe. An example of crosswalks between ETD-MS and Marc-21 (see http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html) is shown in table 1.

<table>
<thead>
<tr>
<th>ETD-MS Element</th>
<th>MARC-21 field &amp; subfield</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.title</td>
<td>245a</td>
<td>A name given in the title page of the resource. e.g. Title of the theses.</td>
</tr>
<tr>
<td>dc.creator</td>
<td>100a</td>
<td>Author/Researcher who is responsible for writing the theses.</td>
</tr>
<tr>
<td>dc.subject</td>
<td>653a</td>
<td>Topic of the content of the theses or keywords.</td>
</tr>
<tr>
<td>thesis.degree.name</td>
<td>502a</td>
<td>Name of the degree for which the theses is submitted. e.g. ‘Masters’ or ‘Doctoral’.</td>
</tr>
</tbody>
</table>

Table: 1

Preprint version of the following paper

OPEN ARCHIVES INITIATIVE- PROTOCOL FOR METADATA HARVESTING (OAI-PMH):

An Open Archive is a computer interface to access a collection of data, where the interface must be OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting) complaint. Current standard OAI is the Protocol for Metadata Harvesting, which describes how two IDR systems can communicate a stream of structured records from one to the other on a periodic basis. An IDR system that contains the source data and that conforms to this protocol is called an Open Archive.

The Open Archives Initiative was launched in October 1999 in an attempt to address interoperability problem among the many existing and independent IDRs. The aim of this Initiative was on high-level communications among IDRs and the simplicity of protocols\textsuperscript{15}. The OAI-PMH is different in comparison to earlier approaches to interoperability. The primary purpose of this protocol is incremental bulk transfer of metadata (harvesting). This protocol does not provide remote searching facility – instead of remote searching, a service provider acquires data from data providers, stores and processes it locally and then provides services to its users\textsuperscript{11}. This protocol helps to make possible metadata available as resource discovery. The OAI-PMH is a request-response protocol that allows one computer to collect metadata incrementally over time from another computer, which is called harvesting\textsuperscript{13}. The protocol is simple to design and suitable for use in various communities in different contexts. To harvest metadata into the union catalogue of ETDs, this project has adopted OAI-PMH interoperability framework. OAI-PMH facilitates an application independent interoperability framework to share metadata among different heterogeneous IDR systems\textsuperscript{4}. The framework is now mainly used for the worldwide consolidation of open access archives that deals with ETDs. Here, the system attempts to design a single window search service for ETDs through harvesting metadata from participating IDRs. It is

Preprint version of the following paper

important to mention here that this single window search service only provides access to metadata – it does not provide the resources themselves. The union catalogue archive of ETDs is designed to function as both service provider and data provider. It harvests metadata from remote IDRs, stores harvested metadata in an internal database and reproduce this metadata through data provider interface. On the other hand this system may also act as data provider for other harvesting tools and systems. There are two participating components in OAI-PMH framework i.e. I) Data Provider; and II) Service Provider.

**DATA PROVIDER:**

Data Providers administer systems that support the OAI-PMH as a means of exposing metadata. A data provider maintains a repository that allows external online access to metadata through OAI-PMH. Any 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.

**SERVICE PROVIDER:**

Service Providers use metadata harvested via OAI-PMH as a basis for building value-added services. A service provider is an entity that harvests metadata from data providers in an attempt to provide higher quality services to users. Once the OPACs (Online Public Access Catalogues) of institutions are made OAI-complaint, the base URLs of such OPACs are to be registered with an OAI-based service provider. The service provider will then harvests metadata from the registered OPACs (Data Provider) on a periodic basis.

There are some advantages of implementing OAI-PMH framework are given below:

I) It provides a new model for scholarly communication in the Internet era. OAI solves the

Preprint version of the following paper

problem of Interoperability;

II) To implement OAI-PMH is easy and simple;

III) It is open; and

IV) It uses current web standards wherever applicable.

The OAI-PMH supports six verbs that are used to send requests to the Data Providers.

Table 2 describes this.

**Response to these OAI-PMH verbs request are given below**

<table>
<thead>
<tr>
<th>Request verb</th>
<th>Response function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify</td>
<td>To retrieve information about description of IDR, e.g. name of the repository, the protocol version, administrator’s email etc.</td>
</tr>
<tr>
<td>ListMetadata Formats</td>
<td>To retrieve the metadata format from the IDR.</td>
</tr>
<tr>
<td>ListSets</td>
<td>To retrieve the sets and subsets from an IDR.</td>
</tr>
<tr>
<td>GetRecord</td>
<td>To retrieve an individual metadata record from an IDR.</td>
</tr>
<tr>
<td>ListIdentifiers</td>
<td>To retrieve unique identifiers from an item.</td>
</tr>
<tr>
<td>ListRecords</td>
<td>To harvest records from an IDR.</td>
</tr>
</tbody>
</table>

**Table: 2**

![Diagram of Data Provider and Service Provider in OAI-PMH Framework](image)

**Functions of Data Provider and Service Provider in OAI-PMH Framework**

Preprint version of the following paper

DESIGNING FRAMEWORK:

Design and development of harvesting framework for ETDs requires following steps, strategies and planning. There are three major components -

i) Development of software framework;
ii) Selection, installation and configuration of harvesting tool; and
iii) Selection of IDRs and collection of essential attributes for harvesting (name of the IDR, resource URL, base URL, etc.).

The prototype harvesting framework named as UniTHES, is based on open source software and open standards. This framework, like many other systems is based on LAMP architecture. The versions of different software are as follows –

- PHP >= 4.2.x (including PHP 5.x); Microsoft IIS requires PHP 5.x
- MySQL >= 3.23.23 (including MySQL 4.x/5.x)
- Apache >= 1.3.2x or >= 2.0.4x or 2.0.5x /Microsoft IIS 5.x or 6.x
- **Operating system (OS):** Any OS that supports the above software, including Linux, BSD, Solaris, Mac OS X, Windows (preferably NT based Windows flavors).

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. These are as follows –

**Group A: LAMP related activities**

Preprint version of the following paper

PKP harvester 2.X is based on AMP architecture. So, we have installed Apache, MySQL and PHP before installation of PKP harvester. Although there is no hard and fast rule, the order of installation sequence of this manual is as follows:

Apache (The Apache httpd server is a powerful, flexible, HTTP/1.1 compliant open source Web server):

- Installation of Apache;
- Testing of Apache; and
- Apache Configuration and Control.

PHP (PHP is an open source server side scripting language):

- Installation of PHP;
- Configuration of PHP.

MySQL (MySQL, the most popular Open Source SQL database, is developed, distributed and supported by MySQL AB):

- Installation of MySQL;
- Initialization of MySQL Server;
- Creation of database, user and manage permission.

Testing of AMP Links through Scripts:

- Testing PHP-Apache Link;
- Testing PHP-MySQL Link.

Group B: Harvester related functions:

Preprint version of the following paper

This group includes two major activities:

i) Installation of PKP harvester;

ii) Configuration of PKP harvester.

iii) The installation process of PKP harvester is quite easy. It requires two sets of information –
(a) login name and password for the administrator and (b) database details (name of the mysql
database, user of database and password of the database user).

The configuration processes are divided into three groups:

a) Site management (configuration of site-specific details, language, crosswalk, plugins and
reading tools);

b) Archives (creation of archives, managing created archives); and

c) Other administrative functions (layout, customization etc.).

**Group C: Repository related activities**

The most important task of the administrator is to set up archive(s) for metadata harvesting. The UniTHES started with thirty OAI-PMH compatible open access IDRs containing electronic theses and dissertations. For selecting IDRs containing ETDs, some pre-determined criteria were under consideration, i.e. IDRs should have ETDs in English language; each IDR have more than 10 records and should have OAI-PMH base URL. In this research project DOAR (Directory of Open Access Repositories)\(^9\) and ROAR (Registry of Open Access Repositories)\(^10\) were consulted for selecting IDRs. The intrinsic attributes of these repositories are as follows –

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Name of the Sponsoring Institution, Country &amp; URL</th>
<th>Name of the IDR</th>
<th>OAI-PMH base URL</th>
<th>No. of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>University of Lethbridge, Canada, <a href="http://www.uleth.ca/">http://www.uleth.ca/</a></td>
<td>DSpace at the University of Lethbridge</td>
<td><a href="http://www.uleth.ca/dspace-oai/request?verb=Identify">http://www.uleth.ca/dspace-oai/request?verb=Identify</a></td>
<td>15</td>
</tr>
</tbody>
</table>

Preprint version of the following paper

<table>
<thead>
<tr>
<th></th>
<th>Institution</th>
<th>Repository</th>
<th>URL</th>
<th>OSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Learning &amp; Information Services, University of Wolverhampton, United Kingdom, <a href="http://www.wlv.ac.uk/">http://www.wlv.ac.uk/</a></td>
<td>Wolverhampton Intellectual Repository and E-theses</td>
<td><a href="http://wlv.openrepository.com/wlv-oai/request?verb=Identify">http://wlv.openrepository.com/wlv-oai/request?verb=Identify</a></td>
<td>63</td>
</tr>
<tr>
<td>5</td>
<td>University of Johannesburg (UJ), South Africa, <a href="http://www.uj.ac.za/">http://www.uj.ac.za/</a></td>
<td>DSpace at the University of Johannesburg</td>
<td><a href="http://ujdigispace.uj.ac.za:8080/dspace-oai/request?verb=Identify">http://ujdigispace.uj.ac.za:8080/dspace-oai/request?verb=Identify</a></td>
<td>53</td>
</tr>
<tr>
<td>7</td>
<td>COM•DTU (Department of Communications, Optics &amp; Materials), DTU (Technical University of Denmark), Denmark, <a href="http://www.dtu.dk/">http://www.dtu.dk/</a></td>
<td>System Competence Area Document Server</td>
<td><a href="http://sysdoc.com.dtu.dk/oai2d.py/?verb=Identify">http://sysdoc.com.dtu.dk/oai2d.py/?verb=Identify</a></td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>Napier University, United Kingdom, <a href="http://www.napier.ac.uk/">http://www.napier.ac.uk/</a></td>
<td>Repository@Napier</td>
<td><a href="http://researchrepository.napier.ac.uk/cgi/oai2?verb=Identify">http://researchrepository.napier.ac.uk/cgi/oai2?verb=Identify</a></td>
<td>48</td>
</tr>
<tr>
<td>13</td>
<td>University of St Andrews, United Kingdom, <a href="http://www.st-andrews.ac.uk/">http://www.st-andrews.ac.uk/</a></td>
<td>DSpace at St Andrews University</td>
<td><a href="http://research-repository.st-andrews.ac.uk/dspace-oai/request?verb=Identify">http://research-repository.st-andrews.ac.uk/dspace-oai/request?verb=Identify</a></td>
<td>429</td>
</tr>
</tbody>
</table>

Preprint version of the following paper

Preprint version of the following paper

<table>
<thead>
<tr>
<th>No.</th>
<th>Institution</th>
<th>Repository Details</th>
<th>URL</th>
<th>OAI-Identify Query</th>
<th>4-digit Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>University of Leicester Library, University of Leicester, United Kingdom, <a href="http://www.le.ac.uk/">http://www.le.ac.uk/</a></td>
<td>Leicester Research Archive</td>
<td><a href="http://lra.le.ac.uk/dspace-oai/request?verb=Identify">http://lra.le.ac.uk/dspace-oai/request?verb=Identify</a></td>
<td></td>
<td>206</td>
</tr>
<tr>
<td>26</td>
<td>Kansas State University, United States, <a href="http://www.k-state.edu/">http://www.k-state.edu/</a></td>
<td>K-State Research Exchange</td>
<td><a href="http://krex.k-state.edu/dspace-oai/request?verb=Identify">http://krex.k-state.edu/dspace-oai/request?verb=Identify</a></td>
<td></td>
<td>1181</td>
</tr>
<tr>
<td>27</td>
<td>Northeastern University, United States, <a href="http://www.northeastern.edu/">http://www.northeastern.edu/</a></td>
<td>IRis</td>
<td><a href="http://iris.lib.neu.edu/iii/oairep/OAIRepository?verb=Identify">http://iris.lib.neu.edu/iii/oairep/OAIRepository?verb=Identify</a></td>
<td></td>
<td>1254</td>
</tr>
<tr>
<td>28</td>
<td>Oregon State University (OSU), United States, <a href="http://oregonstate.edu/">http://oregonstate.edu/</a></td>
<td>DSpace-DEV at Oregon State University</td>
<td><a href="http://ir.library.oregonstate.edu/dspace-oai/request?verb=Identify">http://ir.library.oregonstate.edu/dspace-oai/request?verb=Identify</a></td>
<td></td>
<td>97</td>
</tr>
<tr>
<td>29</td>
<td>SFU Library (Simon Fraser University Library), SFU (Simon Fraser University), Canada, <a href="http://www.sfu.ca/">http://www.sfu.ca/</a></td>
<td>SFU Institutional Repository</td>
<td><a href="http://ir.lib.sfu.ca/dspace-oai/request?verb=Identify">http://ir.lib.sfu.ca/dspace-oai/request?verb=Identify</a></td>
<td></td>
<td>4076</td>
</tr>
<tr>
<td>30</td>
<td>Drexel University, United States, <a href="http://www.drexel.edu/">http://www.drexel.edu/</a></td>
<td>iDEA:Drexel E-repository and Archives</td>
<td><a href="http://idea.library.drexel.edu/dspace-oai/request?verb=Identify">http://idea.library.drexel.edu/dspace-oai/request?verb=Identify</a></td>
<td></td>
<td>466</td>
</tr>
</tbody>
</table>

**Table 2**

UniTHES so far harvested metadata from thirty repositories and agglomerated a total of more than twenty one thousand open access ETDs. It supports localized searching of metadata elements in two modes – simple and advance. Users can limit search in a single repository or a group of repositories (by default a given search session includes all the available repositories). Search can be filtered by DC metadata elements like title, author, date range, language etc. UniTHES can harvest metadata from DCMES (Simple Dublin Core metadata), MARC 21 bibliographic format and ETD-MS (metadata for electronic theses and dissertations).
A snapshot of browsing ETD archive

Preprint version of the following paper

CONCLUSION:

UniTHES is beginning of a new era of localized resource discovery services that can harvest and update metadata from different OAI-PMH compatible open access ETD repositories. At present it includes thirty large-scale open access repositories (based on carefully crafted selection criteria) dedicated to the domain of ETD (in different disciplines). This system has the capability to include new OAI-PMH compatible ETDs and to update records for existing ETDs.

Preprint version of the following paper

for the purpose of developing a comprehensive single window local search service for open access ETD repositories.

References:


Preprint version of the following paper


Preprint version of the following paper