ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2015

Metadata Harvesting From Selected Institutional Digital Repositories in India: A Model to Build a Central Repository

Soumen Teli

Professional Trainee, Department of Central Library, IIT Kharagpur, West Bengal, India

ABSTRACT: This paper tried to discuss the phenomena for designing and developing a central metadata harvesting repository from selected Indian Institutional Digital Repositories. The metadata Standard format used is the Dublin Core Metadata. For metadata harvesting from the IDRs we used the Open Archives Initiative Protocol for Metadata Harvesting standard version 2.0. The study primarily tried to focus on building a central indexed repository. The repository is configured using Dspace version 4.1 and harvesting is done using its user interface. The number of digital repositories is increasing enormously around the world. With reference to DOAR in India there are presently 68 digital repositories. User's want relevant information according to their query with minimum interval of time. Henceforth they compled to search the individual repositories for getting his or her desire documents. This study tries to solve the mentioned problem of "searching individual repositories" by building a central indexed repository and providing a single search dialog box. In order to allow users to search efficiently from various institutional digital repositories we have configure a central indexed repository which will harvest metadata as well as Bitstream only if ORE support is configured in the data provider server. Metadata harvesting is the process of aggregating metadata from various data providers. Then DSpace service uses these aggregated metadata and apache SOLR indexes them and finally provides a resource discovery solution to the library patrons.

KEYWORDS: Repository, Open Archive, Metadata, OAI-PMH, Dublin Core, Open access, Harvesting, OAI-ORE

I. Introduction

The development of open source digital library software brings an avenue in the pattern of library and information services to collect, organise, preserve and disseminate information to the diverse user community. The number of digital repositories and digital publications distributed all over the India is increasing rapidly. It is found that major document deposited in Institutional repositories are Theses, Dissertations, Conference papers, Journal article, Reports, Patents etc. Such growth gives new opportunities. It allows using distributed metadata to create novel network services with content and functionality not possible to achieve before. As the collection grew the need for tools to manage the content exchange local to the federated repository become evident. Data exchange between repositories is a component of Cooperative repository initiatives. The essence of harvesting is to enable access to web-accessible material through inter operable repositories for metadata sharing, publishing and archiving. The sharing of knowledge may lead to further development in to the same discipline or related discipline. OAI-PMH specifies two players in the harvesting process- Data provider, who create structural metadata and expose them for harvesting and Service provider, who harvest and normalize the structured metadata, providing a searchable interface to search for and retrieve metadata records. The harvesting process is consisting of the service provider using HTTP to request information from the data provider, which responds in accordance with the established protocol.

Metadata is structured information that describes, explains, locates or otherwise make it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information.
Metadata harvesting is a technique for gathering together of metadata from a number of distributed repositories into a combined data store.

DOI: 10.15680/IJIRSET.2015.0404018

II. OBJECTIVE OF THE STUDY

The objectives of the study are: To store and preserve other institutional digital assets. To collect content from different repositories into a single location and To provide open access of different Institutional research outputs by self archiving.

III. LITERATURE REVIEW

Review of related search, and search for related literature on the chosen topic of research work is an important step before the actual research work is undertaken. By the literature survey a user can directly refer to that particular document which is relevant to the concept, which help to avoid duplication of previous research work and also help to discover something new or expanding the existing knowledge. The review of literature, in the present study covers macro and micro documents, e-resources as well as a few websites relevant to the topic concerned.

OAI: The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication.

OAI-PMH: The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. Data Provider is repositories that expose metadata via OAI-PMH. Service providers then make OAI-PMH service request to harvest that metadata. OAI-PMH is a set of six verbs of services that are invoked within HTTP.

OAI-ORE: Open Archives Initiative Object Reuse and Exchange (OAI-ORE) defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation.

OAI-PMH Verbs: These are the Verbs defined in the OAI-PMH

GetRecord: This verb is used to retrieve an individual metadata record from a repository.
Identify: This verb is used to retrieve information about a repository.
List Identifiers: This verb is an abbreviated form of List Records retrieving only headers rather than records.
List Metadata Formats: This verb is used to retrieve the metadata formats available from a repository.
ListRecords: This verb is used to harvest records from a repository
ListSets: This verb is used to retrieve the set structure of a repository
Record Format: At the lowest level a data provider must support the simple Dublin Core record format ('oai_dc'). This format is defined by the OAI-PMH DC XML schema. Data providers may also provide metadata records in other formats. Example OAI DC metadata record,
The following are an example of record format

```
<oai_dc:dc>
<dc:title>Electrokinetics in Narrow Fluidic Confinements</dc:title>
<dc:creator>Chakraborty, Jeevanjoyti</dc:creator>
<dc:date>2013</dc:date>
<dc:type>Thesis</dc:type>
<dc:identifier>http://10.17.250.203:8080/xmlui/handle/123456789/3173</dc:identifier>
<dc.subject>Geometry</dc.subject>
</oai_dc:dc>
```

IV. INSTITUTIONAL DIGITAL REPOSITORIES IN INDIA

Data collected from Directory of Open Access Repository (DOAR) shows there are 68 IDRs in Indian subcontinent. Dspace, ePrints and many other open source softwares are being used for creating repositories. The source of the data is openDOAR data collected on 26th Dec 2014. With reference to the above data it may be said as if a librarian selects “ePrints” or “DSpace” for creating IDR then it will be a good decision, because sufficient help is available as many institutions are using one of the two mentioned software.

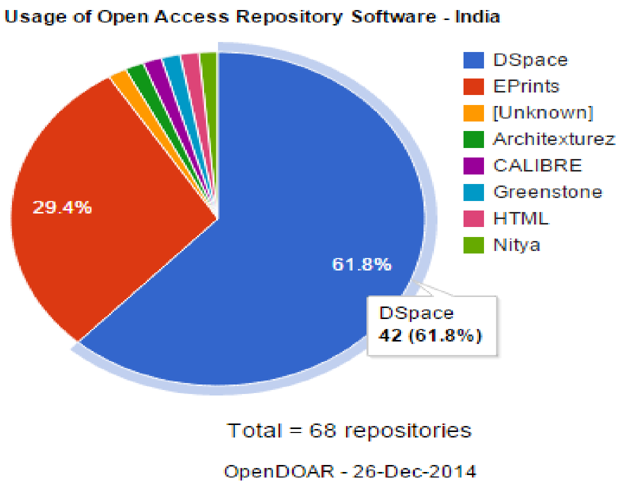


Fig 1

The findings show that 61.8% i.e. 42 number of institutions uses Dspace and 29.4 % of institutions use EPrints software’s. Henceforth it is being decided for the study that we will create the central repository using Dspace software.

V. RESEARCH METHODOLOGY

Methodology has tremendous roll to execute a project successfully. This article is all about the planning for harvesting of digital contents and creation of a Central Indexed Repository. As mentioned above out of the 68 digital repositories we have randomly selected 13 repositories (i.e. 20% of total repositories) for metadata harvesting. All the selected IDRs are preserving digital contents in various subject areas and out of selected 13 repositories only 6 repositories are founded active for metadata harvesting.

S.N	Name of the Institution	Harvesting Status	Content Size	Software Used
1	Indian Institute of Technology, Bombay	YES	15,914	Dspace
2	Indian Institute of Technology, Delhi	YES	3,388	Eprints
3	Indian Institute of Management, Ahmadabad	YES	12,088	Dspace
4	Indian Statistical Institute	YES	5,867	Dspace
5	Inter University Centre for Astronomy and Astrophysics	YES	2,625	Dspace

6	Raman Research Institute	YES	5,728	Dspace
7	University of Delhi	NO	—	—
8	Vidya Prasarak Mandal, India	NO	—	—
9	Documentation Research and Training Centre (DRTC), Indian Statistical Institute, Bangalore Centre (ISI), India	NO	—	—
10	National Chemical Laboratory Pune	NO	—	—
11	Archives of Indian Labour	NO	—	
12	National Centre for Radio Astrophysics	NO		—
13	Indian Institutes of Science Education and Research, Mohali	NO	—	—

VI. Creation of Central Repositories

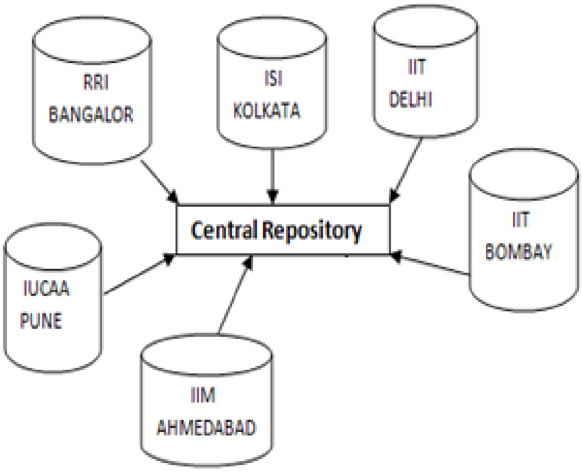


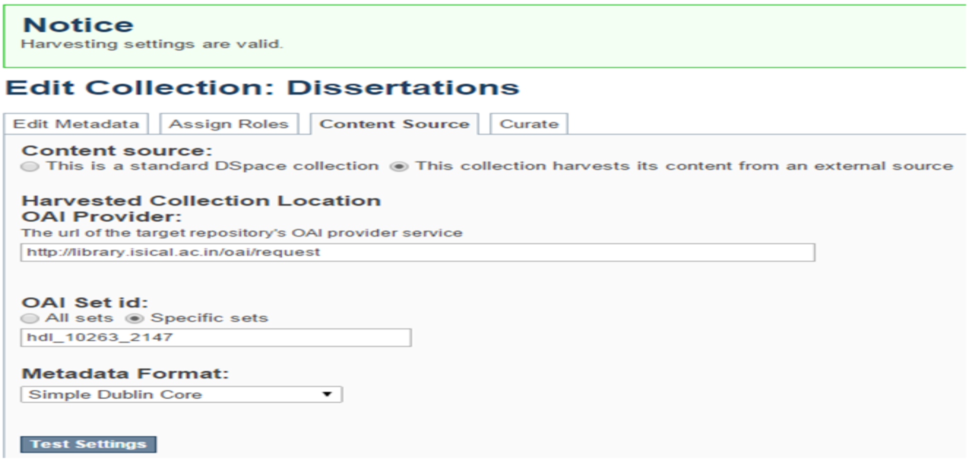
Fig 2

As shown in figure 2 central repository server is build using Dspace version 4.2.It harvests metadata from various IDR on test basis.The single window search facility helps the end user to search all the 7 repositories in one place.This finally helps and benefits the library patrons

VII. Steps of Harvesting

Harvesting is the process accomplished in three steps. First validate the OAI URL, and then decide the set specification to be harvested and finally import the data. This is shown in the screen shots below.

Validity Testing for Metadata Harvesting



Notice
Harvesting settings are valid.

Edit Collection: Dissertations

Content source:
○ This is a standard DSpace collection ● This collection harvests its content from an external source

Harvested Collection Location
OAI Provider:
The url of the target repository's OAI provider service
http://library.isical.ac.in/oai/request

OAI Set id:
○ All sets ● Specific sets
hdl_10263_2147

Metadata Format:
Simple Dublin Core

Fig 3

The important command for harvesting is Identify the OAI Data Provider

URL : http://dspace.library.iitb.ac.in/oai/request?verb=Identify

Output

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-
PMH.xsd">
<responseDate>2014-12-23T14:00:30Z</responseDate>
<request verb="Identify">http://dspace.library.iitb.ac.in/oai/request</request>
<Identify>
<repositoryName>DSpace at IIT Bombay</repositoryName>
<baseURL>http://dspace.library.iitb.ac.in/oai/request</baseURL>
<protocolVersion>2.0</protocolVersion>
<adminEmail>dspace@iitb.ac.in</adminEmail>
<earliestDatestamp>2001-01-01T00:00:00Z</earliestDatestamp>
<deletedRecord>persistent</deletedRecord>
<granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
<compression>gzip</compression>
<compression>deflate</compression>
<description>
<toolkit xmlns="http://oai.dlib.vt.edu/OAI/metadata/toolkit"
xsi:schemaLocation="http://oai.dlib.vt.edu/OAI/metadata/toolkithttp://oai.dlib.vt.edu/OAI/metadata/toolkit.xs
d">
<title>OCLC's OAICat Repository Framework</title>
<author>
<name>Jeffrey A. Young</name>
<email>jyoung@oclc.org</email>
<institution>OCLC</institution>
</author>
<version>1.5.48</version>
<toolkitIcon>http://alcme.oclc.org/oaicat/oaicat_icon.gif</toolkitIcon>
```

```
<URL>http://www.oclc.org/research/software/oai/cat.shtm</URL>
</toolkit>
</description>
</Identify>
</OAI-PMH>
```

The important command for Set Specification is ListSets

URL : http://dspace.library.iitb.ac.in/oai/request?verb=ListSets

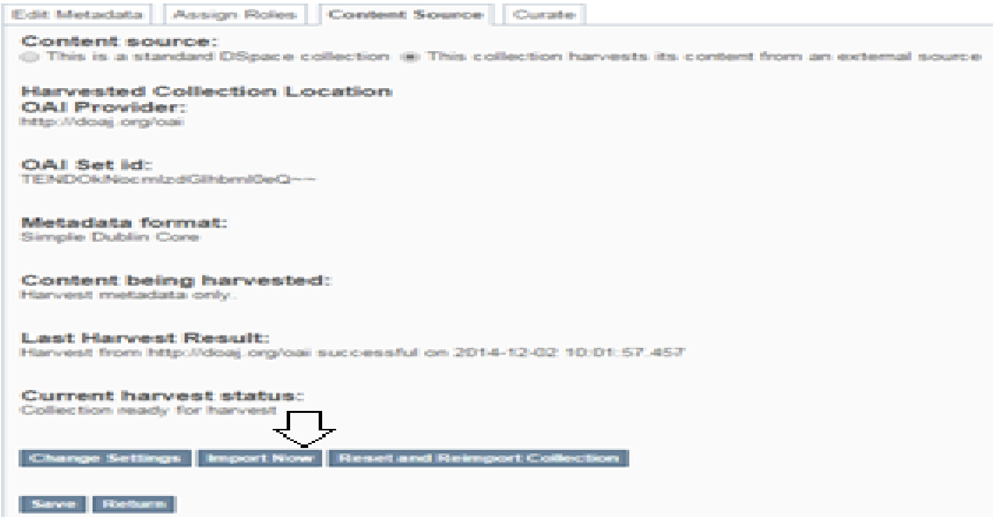
Output

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-
PMH.xsd">
<responseDate>2014-12-23T13:59:44Z</responseDate>
<request verb="ListSets">http://dspace.library.iitb.ac.in/oai/request</request>
<ListSets>
<set>
<setSpec>hdl_100_13681</setSpec>
<setName>Article</setName>
</set>
</ListSets>
```

The OAI gives information on tests an OAI repository must successfully complete in order to be entered in the registry. For example:

- For every protocol request, the repository return a response that is valid XML (the XML successfully passes through an XML parser) and conforms to the XML schema defined for the response (passes the XSV XML schema validator).
- For the ListMetadata Formats request, the repository must return at least one metadata format and in the list of metadata formats, return the mandatory oai_dc metadata format with the URL of the OAI-defined XML schema.

Import Process for metadata import from the data provider



Content source:
○ This is a standard DSpace collection ● This collection harvests its content from an external source

Harvested Collection Location
OAI Provider:
http://doaj.org/oai

OAI Set id:
TENDOkNocmlzdGlhbml0eQ~~

Metadata format:
Simple Dublin Core

Content being harvested:
Harvest metadata only.

Last Harvest Result:
Harvest from http://doaj.org/oai successful on 2014-12-02 10:01:57.457

Current harvest status:
Collection ready for harvest

Fig 4

DOI: 10.15680/IJIRSET.2015.0404018

VIII. HARVESTING STATUS OF IDRS

SL	Institution	Harvesting URL	Metadata Harvested
1	Indian Institute of Technology, Bombay	http://dspace.library.iitb.ac.in/oai/request	9,256
2	Indian Institute of Technology, Delhi	http://eprint.iitd.ac.in/oai/request	3,194
3	Indian Institute of Management, Ahmedabad	http://vslir.iimahd.ernet.in:8080/oai/request	11,034
4	Indian Statistical Institute	http://library.isical.ac.in/oai/request	5,154
5	Inter University Centre for Astronomy and Astrophysics	http://www.iucaa.ernet.in:8080/oai/request	2,566
6	Raman Research Institute	http://dspace.rri.res.in/oai/request	3,297

The above table show that we have harvested around 34,500 metadata from 6 active IDR's. With this we can create a model central repository which will help users to locate and find their required data very easily and in one single place.

IX. CONCLUSIONS

This paper tried to describe a model to building a Central repository. We have illustrated how existing metadata can be harvested fairly easily and using standard tools. It is used to harvest remote IDRs by means of the OAI-PMH protocol. It also gives the user the possibility to search through gathered remote metadata. In fact, any OAI-PMH-enabled repository can be harvested and searched using that service. However, this paper also illustrated the technological side of the harvesting process.

REFERENCES

1. Das , Anup Kumar., Sen, B K. and Dutta, Chaitali "Collection Development in Digital Information Repositories in India"
2. Directory of Open Access Repository URL: http://www.opendoar.org/countrylist.php?cContinent=Asia accessed on 23-12-2014
3. Dublin Core URL: <http://dublincore.org/> accessed on 23-12-2014
4. Frank McCown, Michael L. Nelson, "A Framework for Describing Web Repositories" Proceedings of JCDL pp. 341-344, 2009,
5. Ghosh, S.B. and Das, Anup Kumar "Open access and institutional repositories – a developing country perspective: a case study of India" http://www.ifla.org/IV/ifla72/papers/157-Ghosh_Das-en.pdf (Last accessed on 22.08.2006)
6. OAI "Open Archives Initiative Protocol - Object Exchange and Reuse". Open Archives Initiative, last checked 23-12-2014 http://www.openarchives.org/ore/

7. OAI "ORE User Guide - Primer". Open Archives Initiative, last checked 23-12-2014 http://www.openarchives.org/ore/1.0/primer.html
8. Roy, B. K., Mukhopadhyay, P. and Biswas, S. C. "An analytical study of institutional digital repositories in India". Library Philosophy and Practice 2011.
9. Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., and Smith, M. "The Dspace institutional digital repository system: current functionality". In Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, Houston, Texas, ,p87-97, May 27 - 31, 2003
10. The Open Archives Initiative Protocol for Metadata Harvesting URL: http://www.openarchives.org/OAI/openarchivesprotocol.html accessed on 23-12-2014
11. Van de Sompel, H., Nelson, M.L., Lagoze, C., and Warner,S., "Resource Harvesting within the OAI-PMH Framework", D-Lib Magazine, vol.10, no 12, December 2004,
12. Warner, S. "Exposing and Harvesting Metadata Using the OAI Metadata" June 2001
13. Witt, M. "Object Reuse and Exchange (OAI-ORE)" Library Technology Reports, Vol. 46, No. 4, 2010