Federated Search Service for OAI-compliant Open-Access Repositories in India

Francis Jayakanth[1] and Filbert Minj[2]

[1]JRD Tata Memorial Library
Indian Institute of Science
Bengaluru -560 012
franc@library.iisc.ernet.in

[2]Super Computer Education and Research Centre
Indian Institute of Science
Bengaluru -560 012
filbert@serc..iisc.ernet.in

Abstract. Many of the research institutions and universities across the world are facilitating open-access (OA) to their intellectual outputs through their respective OA institutional repositories (IRs) or through the centralized subject-based repositories. The registry of open access repositories (ROAR) lists more than 2850 such repositories across the world. The awareness about the benefits of OA to scholarly literature and OA publishing is picking up in India, too. As per the ROAR statistics, to date, there are more than 90 OA repositories in the country. India is doing particularly well in publishing open-access journals (OAJ). As per the directory of open-access journals (DOAJ), to date, India with 390 OAJs, is ranked 5th in the world in terms of numbers of OAJs being published.

Much of the research done in India is reported in the journals published from India. These journals have limited readership and many of them are not being indexed by Web of Science, Scopus or other leading international abstracting and indexing databases. Consequently, research done in the country gets hidden not only from the fellow countrymen, but also from the international community. This situation can be easily overcome if all the researchers facilitate OA to their publications.

One of the easiest ways to facilitate OA to scientific literature is through the institutional repositories. If every research institution and university in India set up an open-access IR and ensure that copies of the final accepted versions of all the research publications are uploaded in the IRs, then the research done in India will get far better visibility. The federation of metadata from all the distributed, interoperable OA repositories in the country will serve as a window to the research done across the country.

Federation of metadata from the distributed OAI-compliant repositories can be easily achieved by setting up harvesting software like the PKP Harvester. In this paper, we share our experience in setting up a prototype metadata harvesting service using the PKP harvesting software for the OAI-compliant repositories in India.

Keywords: Open-access, Interoperability, Metadata Harvesting, OAI-PMH, OAI-compliance

1 Introduction

For the further development of knowledge, scholars need to have access to the relevant scholarly literature. However, because of the runaway cost increases of many scholarly journals, even the well-funded libraries in the world can not afford to subscribe to all the scholarly literature that their scholars would like to have access to. This creates an access-barrier to the scholarly literature. This in turn affects the visibility and the potential impact of the scholarly literature. The proponents of OA movement to scholarly literature believe that the access-barrier to the scholarly literature can be overcome if the researchers take advantage of the developments that are taking place in the information and communication technologies by facilitating OA peer-reviewed scholarly literature that they publish.

To achieve open access to peer-reviewed scholarly literature, the Budapest Open Access Initiative has recommended the following two complimentary strategies (http://www.soros.org/openaccess/read):

1. Self-Archiving: Self-archiving is a process wherein, upon the acceptance of a paper for publication, authors deposit a copy of their final-accepted versions f the publications in an Institutional Repository or in a subject-based central repository.

2. Open-access Journals (OAJ): Open access journals are scholarly journals that are available to the reader without financial or other barrier other than access to the internet itself.

Today, there are several free and open source software (FOSS) that could be implemented to facilitate OA to scholarly literature through the IRs.. DSpace, GNU Eprints.org, FedoraCommans are examples of such software. To publish OAJs, one could use Open Journal Systems (OJS) or Topaz or any other journal management software.

The registry of open access repositories (ROAR) lists more than 2850 OA repositories across the world. The share of OA repositories from India as reflected in the ROAR as on date is 90. Not all the 90 repositories are institutional repositories, only 40 of them are. The remaining repositories include subject-based repositories, electronic theses repositories and OA journals. All the 90 repositories in India are using FOSS for maintaining their repositories and all these repositories are OAI-compliant. OAI-compliance is a means through which the individual repositories (data providers) can expose the metadata contained in their repositories. OAI service providers can harvest the metadata from the repositories and build services like searching and browsing on the harvested metadata.

The Cross Archive Search Services for OA repositories in India (CASSIR) is an attempt to build and maintain metadata harvesting service for all the OA institutional repositories from India. It has been developed using the Open Harvester System (OHS), a FOSS metadata harvesting system developed by the Public Knowledge Project (PKP),

2 The Need for Federated Search

Federated Search is the process of performing a simultaneous search of multiple, diverse, and distributed sources from a single search page, with the federated search engine acting as intermediary **(**Lederman, 2010**).** Two possible approaches to facilitate federated search are cross searching the distributed repositories based on a protocol such as Z39.50, or else harvesting of metadata from the distributed repositories into one central repository. The federated search through CASSIR is based on metadata harvesting.

Federated search offers several benefits to the end users including the following:

- One-stop access to multiple information sources where by users save on time and money, and improving the utilization of information sources.
- End users need not be familiar with different search interfaces of individual repositories.
- It serves as a discovery tool for the end users as they are not expected to be aware of the existence of different distributed repositories

3 Federation Through OAI-PMH

Metadata is 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 (NISO, 2004).

Describing a resource with metadata allows it to be understood by both humans and machines in ways that promote interoperability. 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. Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly (NISO, 2004).

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting (Lagoze, Sompel, Nelson, & Warner, 2002). The OAI-PMH gives a simple technical option for data providers to make their metadata available to services, based on the open standards, HTTP (Hypertext Transport Protocol) and XML. Thus, metadata from many sources can be gathered together in one database, and service can be provided based on this centrally harvested or aggregated data. Figure 1 depicts this concept.

Figure 1. OAI-based Service Providers Harvest Metadata from Data Providers

4 Metadata Harvesting From Distributed Repositories

OAI-based service providers harvest metadata from registered OAI-compliant systems and build a central index on the harvested metadata. This central index serves as a discovery tool for end-users, who need not be aware of the existence of distributed repositories. OAIster is one of the earliest OAI-based service providers. It is a union catalog of millions of records representing open access resources that has been built by harvesting from open access collections worldwide using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Today, OAIster includes more than 25 million records representing digital resources from more than 1,100 contributors. Similar union catalogs can be built for specific subject domains or geographical provinces or specific publication types.

CASSIR is a metadata harvesting service for the OA institutional repositories in India.. Through CASSIR, an attempt is being made to build a union catalogue of all the records that have been uploaded in the various OA repositories in India. We have used Open Harvester System (OHS), a FOSS metadata harvesting system developed by the Public Knowledge Project to harvest the metadata. To date, more than 1,80,000 metadata records from about 30+ OA repositories have been harvested.

5 Cross Archive Search Service For Indian Repositories (CASSIR)

Several prototype cross search services based on metadata harvesting have already been established in the country (Hirwande & Hirwande, 2006) and (Hirwade & Bherwani, 2011). However, all the existing federated search services in India are discipline-specific. For example, the Search Digital Libraries (SDL), http://drtc.isibang.ac.in/sdl/, is a federated search service related to the Library and Information Science discipline. Search Engine for Engineering Digital-Repositories (SEED), http://eprint.iitd.ac.in/seed/, is a federated search service for the engineering discipline. A comprehensive federated search service for all the OA repositories in the country has not been attempted as yet. CASSIR (http://smart.ncsi.iisc.ernet.in/oai) is the first such attempt. Figure 2 shows the screen-shot of CASSIR homepage. Such a federated search service offers several benefits to the country in general and researchers in particular. The benefits include enhancement of the visibility of research, avoidance of duplication of research, single point of access to all the research output of the country, and it can also serve as an information management system for the country.

The benefits of federated search service are briefly explained below.

5.1 Enhancing the Visibility

The dissemination of research results and findings is an integral part of both the research process and the career in academia. Researchers would also like their research papers to be read and cited by their peers.

Indian scientists face two problems common to scientists everywhere, but acutely felt by scientists in poorer countries: Access and Visibility (Arunachalam, 2008). They find it difficult to access what has already been published because of the high costs involved in subscribing to journals and databases. Researchers in the rest of the world are unable to access what Indian researchers are doing. This is because much of the research work done in India is published in the national journals, which are not indexed by the leading abstracting and indexing databases. This leads to low visibility and low readership of research papers published by scientists from India.

The visibility problem can be easily overcome by adapting the OA strategies – Publishing in OA journals and/or self-archiving copies of accepted papers either in a subject-based central repository or in an institutional repository.

The federated search for all the OAI-compliant repositories in the country will further aid in enhancing the visibility of research publications by facilitating single point access to all the research publications from the country.

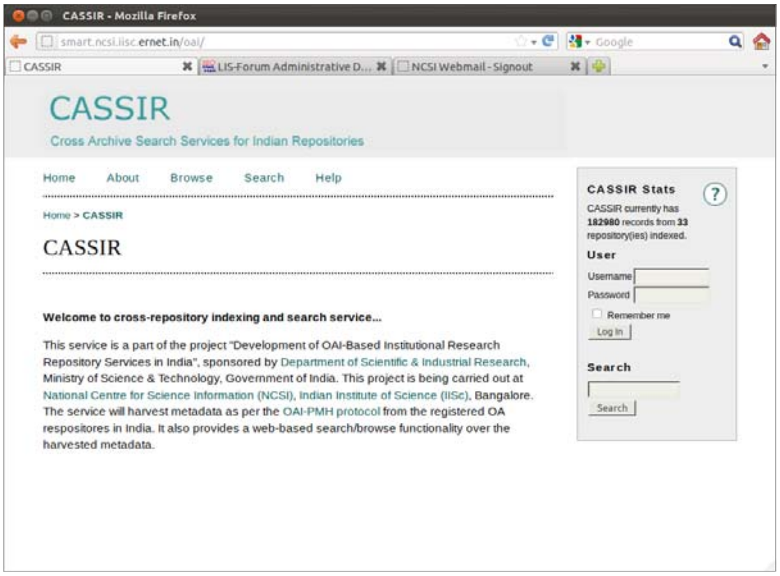
5.2 Aid in Avoiding Duplication of Research Projects

The outcome of research projects among other things will include research papers published in scholarly journals and/or conferences. When such papers are published in the local journals, most of which are not being indexed by leading abstracting and citation databases, fellow researchers and scientists can easily overlook such research publications. This can lead to duplication of research projects. On the contrary, irrespective of the journals or conferences in which the papers are formally published, if the final accepted versions of such papers are uploaded in the OA institutional repositories then the discover-ability of such papers will be much better. This in turn will help in avoiding duplication of research projects.

5.3 Information Management System (IMS) for the Country.

Maintaining an accurate IMS for research publications of a country is a challenging task. For this to be realized every research institution will have to maintain its up-to-date publications' database or there could be a central database where copies of all the research publications of a country are deposited. To the best of our knowledge, to date, no attempt has been made to create a comprehensive publications' database for the country.

If all the universities and research institutions set up their respective OA IRs and populate the IRs with their publications then the creation of federated system from the distributed repositories can also serve as an IMS for the research publications of the country.



CASSIR

Cross Archive Search Services for Indian Repositories

CASSIR

Welcome to cross-repository indexing and search service...

This service is a part of the project "Development of OAI-Based Institutional Research Repository Services in India", sponsored by Department of Scientific & Industrial Research, Ministry of Science & Technology, Government of India. This project is being carried out at National Centre for Science Information (NCSI), Indian Institute of Science (IISc), Bangalore. The service will harvest metadata as per the OAI-PMH protocol from the registered OA respositores in India. It also provides a web-based search/browse functionality over the harvested metadata.

CASSIR Stats

CASSIR currently has 182980 records from 33 repository(ies) indexed.

Figure2. CASSIR Homepage

For the present, CASSIR harvests records from 33 OA repositories in the country. The remaining repositories will be included in due course of time. Table 1 gives the details of the repositories included in the CASSIR that have more than four thousand records.

Repository Name	Institution	Repository URI	No. of Records
Publications of IAS Fellows	Indian Academy of Sciences (IAS)	http://repository.ias.ac.in/index.html	82111
ePrints@IISc	Indian Institute of Science	http://eprints.iisc.ernet.in	33347
NOPR	CSIR-NISCAIR	http://nopr.niscair.res.in/	12533
Eprints@CMFRI	Central Marine Fisheries Research Institute	http://eprints.cmfri.org.in/	7288
Dhananjayarao Gadigil Library	Gokhale Institute of Politics and Economics	http://library.gipe.ac.in/jspui/	5576
OA Digital Repository of IIAP	Indian Inst of Astrophysics (IIAP)	http://prints.iiap.res.in/	5533
NAL IR	National Aerospace Lab.	http://nal-ir.nal.res.in/	4539
RRI Digital Repository	Raman Research Institute (	http://dspace.rri.res.in/	4243
Eprints@NML	National Metallurgical Lab.	http://eprints.nmlindia.org/	4442
Dhananjayarao Gadigil Library	Gokhale Institute of Politics and Economics	http://library.gipe.ac.in/jspui/	5576
DRS@NIO	Nat. Inst of Oceanography	http://drs.nio.org/	4015

Table 1 Sample Listing of Data Providers for CASSIR

6 Issues in Harvesting

There are some software-related issues in harvesting metadata records from certain repositories. For example, the OA digital repository of Manipal University (MU) (http://eprints.manipal.edu), to date, has 3000+ records in the repository. However, CASSIR is able to harvest only 290+ records. Repeated harvesting from this repository each time after flushing the existing records, is not improving the number of records being harvested from the MU repository. However, by using the ListRecords verb from the OAI interface of MU (http://eprints.manipal.edu/cgi/oai2?verb=ListRecords&metadataPrefix=oai_dc) directly, all the records can be listed. Through the OHS, only 290+ records are being harvested.

The problem could be because the OHS software in not able to handle certain special characters that could have crept into the MU repository records. Special characters can easily creep into the records in a repository especially when the metadata details are directly copied from the source (PDF, HTML. Doc, etc) and pasted onto the record being uploaded.

To avoid the possibility of special characters getting embedded, metadata details should be copied from the source and be pasted first on to a text editor. If the metadata contains special characters, such characters can be easily distinguished in a text editor and they should replaced with corresponding ASCII characters. The metadata details should then be copied from the text editor and be pasted on to the record submission screen.

Conclusions

The wide visibility of peer-reviewed scholarly literature is absolutely essential to garner their potential impact. This can be achieved by uploading the copies of the final-accepted versions of all the peer-reviewed scholarly literature in the institutional repositories. Libraries should take the lead in setting up institutional repositories and OA publishing platforms. They should also educate the researchers about the benefits of facilitating OA to their publications and clarify the misconception about OA publishing that some researchers may have.

For developing countries like India, where good number of research papers are being published in the local journals, facilitating OA to its research publications will enhance their visibility. A centralized harvesting service like the CASSIR will serve a window to the research activities in the country. Such a service can showcase the research activities of the country to the rest of world, can help in avoiding duplication of research projects, and can aid in identifying the research areas that needs a thrust.

Acknowledgment

The work reported in the paper was carried out as part of the project, Development of Institutional Research Repository Services in India", funded by the Department of Scientific and Industrial Research (DSIR), Government of India. The authors would like to thank the DSIR for the support.

References

Arunachalam. S. (2008). Open Access in India: Hopes and Frustrations. Proceedings ELPUB 2008 Conference on Electronic Publishing - Toronto, Canada - June 2008. Retrieved on May10, 2012 from: http://elpub.scix.net/data/works/att/271_elpub2008.content.pdf.

Hirwade, M.A. & Bherwani, M.T. (2011). Metadata Harvesting: Tools and Services in India. SRELS Journal of Information Management, 48(4), 389-398.

Lagoze, C., Somplel, Van De, H., Nelson, Michael, & Warner, Simeon. (2002). The Open Archives Initiative Protocol for Metadata Harvesting – Version 2.0. Retrieved May 10, 2012 from http.//www.openarchives.org/OAI_protocol/openarchivesprotocol.html.

Mangala, Hirwade, & Ani, Hirwade. (2006). Metadata Harvesting Services in India, Library Herald, 44(4), 275-282.

NISO (National Information Standards Organization). (2004). Understanding Metadata. Bathesda, MD, USA: NISO Press. Retrieved 10 May, 2012 from http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

Sol, Lederman. (2010). Federated Search Primer. Retrieved May 10, 2010 from http://www.deepwebtech.com/wp-content/uploads/2010/12/Federated-Search-Primer-1.pdf.