HARVESTERS AND FEDERATED SEARCH ENGINES

Baban Kumbhar
Librarian
Rayat Shikshan Sanstha’s
Dada Patil Mahavidyalaya, Karjat,
Dist Ahmednagar
E-mail:kumbharbn@gmail.com

1. Introduction:
The Open Archives Initiative (OAI) develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. It is dedicated to solving problems of digital library interoperability. Its focus has been on defining simple protocol, most recently for the exchange of metadata from archives. Open Archives initiative has its root in an effort to enhance access to print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the open Archive program. The fundamental technological framework and standards that are developing to support this work are however, independent of both the type of content and promise to have much broader relevance in opening up access to a range of digital materials.

The OAI evolved out of a need to increase access to scholarly publication by supporting the creations of inter operable digital libraries. As a first step towards such interoperability a metadata harvesting protocol was developed to support the streaming of metadata from one repository to another. The ultimately to a provider of user services such as browsing searching. The name OAI means open means the protocol is openly documented and metadata is exposed to least some peer group. Archives mean collection of stuff. The OAI uses the term Archives in broader sense: As a repository for the stored information. Initiative means that OAI is happening at break-neck speed.

2. Open Archives Initiative Protocol for Metadata Harvesting (OAI – PMH):
OAI-PMH (Open Archives Initiative Protocol for metadata Harvesting) is a protocol developed by the open Archive Initiatives. It is used to harvest the metadata description of the records in an archives so that service can be built using metadata from many archives OAI-PMH is based on client server architecture in which “harvesters” request information on updated records from repositories. Request for data can be based on date stamp range, and can be restricted to named sets defined by the provider. Data provider (repositories) is required to provide XML metadata in Dublin core format and may also provide it in other XML formats.
3. OAI - PMH Concepts:

**Harvester:** a client application that make an OAI – PMH request.

**Repository:** Network accessible server able to process a OAI – PMH request.

**Resource:** The stuff the metadata is about.

**Item:** A constituent of repository, conceptually, it is the container of the metadata.

**Identifier:** Unique Identifier that unambiguously identifies an item within a repository.

**Record:** An XML – enclosed set of Metadata expressed in a specific format.

**Datestamp:** Date of creation/ modification/ deletion of record.

**Set:** Optional construction for grouping items in the purpose of selective harvesting.

4. OAI – PMH Request:

The OAI – PMH is based on HTTP (*Hyper Text Transfer Protocol*) Request arguments are issued as GET/ POST parameters, OAI – PMH support six request types ( known as “verbs”)

**Identify** : Retrieve Repository information.

**List of Metadata formats:** What metadata formats in repository?

**List sets** : Retrieve repository set structure.

**Get Record** : Retrieve a single metadata record.

**List Records** : Harvest records from a repository.

**List Identified** : Harvest record headers only.

5. OAI – PMH Responses:

Response is encoded in XML syntax. OAI – PMH supports any metadata format enclosed in XML. Dublin core in the minimal format specified for basic interoperability.

- General Information
- Metadata formats
- Set structure
- Record Identifier
- Metadata.

6. Harvest Pattern:

The search on harvest pattern uses an intermediate repository to support searching repositories that expose a harvesting interface. A registry maintains a list of repositories that can be harvested. A harvester harvests all metadata contained in the repositories and stores them in a centralized repository that can be searched by search clients.

7. Harvesters:

1. PKP Harvester.
2. ODL Harvester
3. UIUC Java Harvester
4. UIUC VB Harvester
5. I Via Harvester
6. My OAI Harvester
7. CPAN OAI Harvester
8. Virginia Tech Perl Harvester

1) PKP Harvester:
PKP (Public Knowledge Project) was developed at the University of British Columbia for OAI metadata harvesting and retrieval. The entire site including harvester and search engine works out of the box. The site install easily into a LAMP based server (Linux-Apache-MySQL – PHP) without writing configuration files / installing non LAMP dependencies. The code is readable and well documented. This software is flexible. So it can be installed on windows using open source tools alone.

1.1 Harvesting in PKP:
Harvesting in PKP is managed through a web page from a server running PHP and MySQL. In the browser adding an archive and running it is simple. In test run over 8000 metadata records from the DLESE database were found and indexed.

1.2 PKP search Engine:
The PKP search engine showed varied results. A search for “high school” did not always surface the words “high” and “School” in the metadata record. This appears to be an “or” search by default. Over 4000 of the 8000 records were shown to contain at least one of these search terms. Using “high school course” reveals the same result. A quick scan of the linked record showed to evidence of the word “course”,

1.3 GUI:
Among GUI interfaces, this is an impressive design. It is easy to navigate the site and find the administrative tools to harvest archives and search them. Links lead to an advanced search. Administration access is password protected. With administration access, archives can be searched. Without administrative access the user can browse the retrieved records.

8. Metadata Harvesting:
The common definition of metadata is “data about data” Metadata is one of the important concepts for the description, organization exchange and retrieval of information in the network environment when cranny metadata records to describe a resource, thesauri and
controlled vocabularies are widely used as a basis for both resources description, discovery and information management purposes. The minimum metadata sets are required for harvesting such as document Identifier Title, description and subjects OAI – PMH is mechanism for harvesting XML formatted Metadata from distributed collections of metadata. OAI – PMH provides the basic information discovery environment that relies on transferring metadata on masse from one server to another in a network of information systems based on the open standards HTTP and XML.

The metadata that is harvested may be in any format that is agreed by a community ( or by any discrete set of data and service providers), although unqualified Dublin core is specified to provide a basic level of interoperability, Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested “aggregated” data. The link between this metadata and related content is not defined by the OAI protocol. It is important to realize that OAI – PMH does not provide a search across this data. It simply makes it possible to bring the data together in one place.

9. Metadata Harvesting Services in India:

1. CASSIR: Cross Archive Search Service for Indian Repositories.
2. Open Index Initiative (OII)
3. Open J- Gate
4. Scientific Journal Publishing in India : Indexing and Online Management (SJPI)
5. SDL: Search Digital Libraries.

1. CASSIR:

Cross Archive search service for Indian Repositories. The CASSIR is a metadata harvesting service for Indian repositories. This project initiated by National centre for Service communication (NCSI) of Indian Institute of Science (IISC) and supported by DSIR. This web-based search and browse service is a part of the ongoing project entitled ‘Development of OAI – based Institutional Research Repository services in India: In this service OAI – PMH compliant software i.e. PKP harvester, harvest metadata from registered open access repositories in India. CASSIR serves a functionality of cross searching of repositories.

2. Open Index Initiative (OII):

The open Index initiative ( OII) is a collaborative efforts 16 + volunteers working in Indian social science libraries across 5+ Indian cities, with a basic objective to develop an exhaustive online database with index and abstract to journal articles, book reviews, conference papers, and working papers published in Indian social science journals /
published by Indian social science institutions. The open Index Initiative (OII) portal facilitates information on:

- Form coming conferences/ workshops,
- Search Book reviews
- Index to Journal/ Articles.
- online Directory of social science Institutions
- Theses and dissertations
- Union catalogue of Journals.

**Open Index Initiatives**

OII is a public access database that promotes scholarly literature in social sciences. OII system gets a wide variety of data sets, which is unparalleled to the metadata harvesting services established elsewhere.

**3. Open J-Gate:**

Open – J – Gate is launched by Informatics India Ltd. on Feb. 27, 2006. It is a searchable portal of open access journals. The bibliographic database is maintained in a well-indexed database, having the metadata such as, title, author, email – ID, source, abstract and full text links to each article. It has quick search, advanced search, and browse by journal option. It has quick search, Advanced search, and browse by journal options, when search results are displayed each retrieved record provides details of the articles, and link to full text article of publisher site.
4. **Scientific Journal Publishing in India: Indexing and online Management (SJPI):**

The SJPI cross Journal search service is a part of the SJPI project. In this service OAI – PMH compliant software PKP Harvester, harvests metadata from the sample full text contents of participating Indian scholarly journals.

5. **SDL: Search Digital Libraries:**

The search Digital, Libraries (SDL) is a metadata harvesting service for open access repositories in the area of library and information science. This project initiated by the Documentation Research and Training centre (DRTC) of the Indian statistical Institute (ISI Bangalore) In this service, An OAI – PMH compliant software PKP (Harvester) Harvests metadata from registered open access repositories in the subject area of library and information science.

At present, SDL covers nine open access repositories including one from India.

2. Australian Library and Information Association e – print
3. CALTECHLIB
4. CCSD : Science del Information et de la communication, France,
5. DLIST University of Arizona
8. Librarian’s Digital Library (LDL)
9. OCLC Research Publication.
10. University of North Carolina, USA.

SDL serves a functionality of cross- searching subject specific repositories.

10. **Federated Search Engine:**

Federated searching that is known as meta-searching or cross database searching that provides facility to user to search many networked information resources from one interface. Using federated search engine is an attempt to improve the accuracy and relevance of individual searches while reducing the amount of time required searching specific resources one by one. In addition, by being able to search many sources simultaneously, more content will become visible to the user quickly; Federated search engine uses a distributed search method across heterogeneous databases using multiple searching protocols. “Some specialized federated search engines are limited to metadata harvesting, searching homogenous repositories, or using multiple searching protocols. “Some specialized federated search engines are limited to metadata harvesting, searching homogenous repositories, or
using a limited value number of protocols. Because of the special nature of these applications, they have limited value for general library purposes."

11. **Features of Federated search Engine:**
   1. Ability to search multiple databases concurrently,
   2. Search databases in real time
   3. Simple and advanced search capabilities.
   4. De-duplication of records – removing duplicates according to a set of rules (e.g. which database takes precedence)
   5. Merged search results.
   8. Save, Print, and e-mail results.
   9. Exporting of results to a file.
   10. Extensive user authentication support.
   11. Display database by categories.
   12. Compliance with open URL resolve
   14. Ability to search local and remote collections as well as internet resources and search engines using three models http (screen scrapping), 239.50 and XML.
   15. Personated access to resources.
   16. Ability to access to electronically available content without further authentication.
   17. Compliant with guidelines for accessibility by people with disabilities.
   18. Relevance ranking.
   19. Unlimited simultaneous users.
   20. Ability to link into interlibrary loan (ILL) system.
   22. Ability to be branded to match the look and feel associated with local organization.
   23. Ability to distributed federated research feature can be built into various sections of the organization web site.
ARIADE FEDERATED SEARCH ARCHITECTURE

CONCLUSION:
The search on harvesting uses an intermediate repository to support searching repositories that expose a harvesting interface. A registry maintains a list of repositories that can be harvested. A harvesting harvests all metadata contained repositories and stores them in a centralized repository that can be searched by search clients. PKP (Public Knowledge project) open Archives Harvest handles the mass of data well, both downloading and presenting metadata. Metadata Harvesting services in India are CASSIR, OII, Open J-Gate, SJPI, and SDL are explained.

Federated search engine able to search many sources / databases simultaneously. Federated search engine should not wait for the completion of all search operations. Benefit of federated. Searching is one-stop shopping. Federate searching includes. Resource discovery, wrapper induction, resource representation, resource selection, result merging federated
search engines ability to search multiple repositories without having to learn specific seven options for each repository.

References:


3) Warne, S. Exposing and harvesting metadata using the OAI metadata harvesting protocol, a tutorial Accessed on 20/01/2011
library.web.cern.ch/library/Webzine/4/papers/3

4) Analysis Report on Federated Infrastructure and Application profile
http://www.ariadne –eu.org/

rcdl.ru/doc/2004/paper34.pdf

www.diglib.org/aquifer/oct2504/harvesting.pdf

ir.inflibnet.ac.in/dxml/bitstream/handle/1944/352/04cali_44.pdf?

www. infotoday.com/i/oct 03/ hunel. shtml.

9) Wahi J.H.Introduction to federated search accessed on 25/09/2014