OAI-PMH: A TOOL FOR METADATA HARVESTING AND FEDERATED SEARCH

Dipen Deka

Abstract

This paper focuses on the importance of OAI-PMH in the aspect of accessibility to the digital repositories. The basic structure of OAI-PMH and its functional elements are given along with some existing metadata harvester services of India. The paper discusses about the PKP Harvester software and its users. Concludes that OAI-PMH is an effective solution of the problem of lack of interoperability.

Keywords : Digital repositories, federated search, interoperability, OAI-PMH, Metadata Harvesting

1. Introduction

After the rapid growth of digitization activities the number of digital repositories of various educational, research institutions is also increasing. But no one can say that the digitization is the answer to the problem of proper access of information. The accessibility of these vast and diverse resources is a very difficult task. The lack of interoperability is one of the most significant problems that digital repositories are facing today. In general interoperability is the ability of systems, organizations and individuals to work together towards common or diverse goals. In the technical arena it is supported by open standards for communication between systems and for description of resources and collections, among others. According to Priscilla Caplan search interoperability is 'the ability to perform a search over diverse set of metadata records and obtain meaningful results'. Interoperability is a broad term, touching many diverse aspects of archive initiatives, including their metadata formats, their underlying architecture, their openness to the creation of third-party digital library services, their integration with the established mechanism of scholarly communication, their usability in a cross-disciplinary context, their ability to contribute to a collective metrics system for usage and citation, etc. [3]. Even the powerful search engines have failed to make index of the resources. One of the solutions of this problem is federated search. It is a multiple searching of online databases at the same time to give the users useful results. The traditional search engines use crawler technology and as a result a large volume of databases remained unseen, but the federated searching solves this problem and makes this deep web searchable without visiting them individually. Interoperability is considered here primarily in the context of resource discovery and access.

2. OAI PMH

The evolution of OAI-PMH is one of the solutions to overcome the problem of lack of interoperability. The Open Archives Initiative- Protocol for Metadata Harvesting (OAI-PMH) was designed to facilitate the technical interoperability among distributed digital repositories and archives. It provides an application independent interoperability framework based on metadata harvesting that can be used by a variety of communities who are engaged in publishing content on the Web [2]. The objective of OAI-PMH is to develop a low-barrier, lightweight framework to facilitate the information discovery of content in distributed archives. OAI-PMH has been a success to a great extent, and it has speeded the development of federated service providers such as Arc and OAIster.

2.1 Component of OAI PMH

There are two main components of OAI-PMH. These are as follows

- Service Provider
- Data Provider

2.1.1 Service provider

Service Providers use metadata harvested via the OAI-PMH as a basis for building value-added services [1]. They are entities that harvest metadata from Data providers in order to provide higher-level service to users. Their job is similar to the web-crawlers of the Internet search engines. They go to the individual repositories to harvest their entire metadata, collects in its database in the XML format. The collected metadata is then parsed to provide an integrated search interface and browsing indices to the collections of all the participating data providers/repositories. Service Providers, or harvesters, use metadata harvester via the OAI-PMH as a basis for building value-added services, such as building subject gateways, email alerts, etc. Some of the popular service providers are as follows

OAIster: OAIster is a project of the University of Michigan Digital Library Production services. Its goal is to create a collection of freely available, difficult-to-access, academically-oriented digital resources that are easily searchable by anyone.

NCSTRL: The Networked Computer Science Technical Reference Library (NCSTRL) is an international collection of computer science research reports made available for non-commercial use from over 100 participating organizations worldwide. The organizations that participate in NCSTRL include Ph.D. granting computer science departments, research laboratories, E-Print repositories, and electronic journals. The documents in NCSTRL are almost all textual, ranging in size from 100-plus page doctoral dissertations to short technical reports.

METALIS: It is a Service Provider for the Library and Information Science. It harvests metadata from institutions that offer full-text papers and documents in library and information science. It harvests from the data providers like ArXiv, LDL of DRTC, E-Prints in LIS, Digital Library of Information Science and Technology, CNR Bologna Research Library etc.

2.2.2 Data Provider

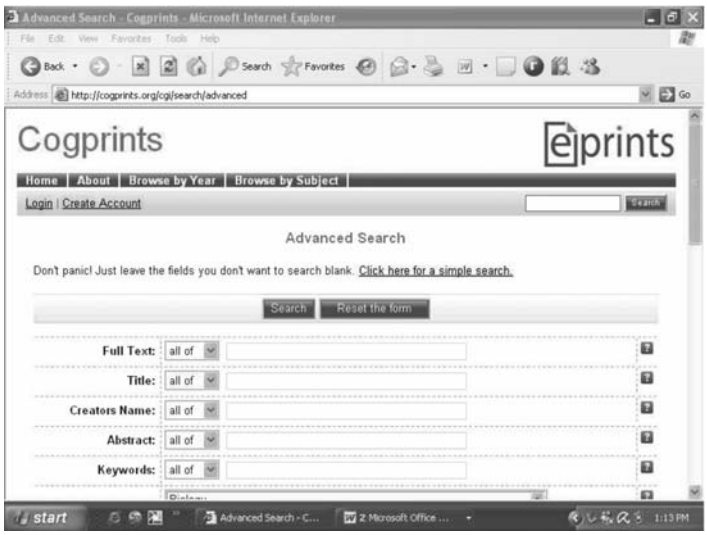
Data Providers administer systems that support the OAI-PMH as a means of exposing metadata [1]. Data providers refer to repositories or archive of a digital content with some kind of metadata describing the content and are willing to share metadata with others via well-defined OAI protocols. The Data Providers expose their metadata, by installing a piece of software, in such a manner that harvesters can harvest their metadata to build value added services. Data Providers, or repositories, administer systems that support the OAI-PMH as a means for exposing their metadata. Here *data* means any kind of digital content, including text, images, sound, and multimedia. Some Data Providers are listed below:

ArXiv: ArXiv is an e-print service in the fields of physics, mathematics, non-linear science and computer science. The contents of ArXiv conform to Cornell University academic standards. ArXiv is owned, operated and funded by Cornell University, a private not-for-profit educational institution. ArXiv is also partially funded by the National Science Foundation.

E-Prints in Library and Information Science (E-LIS)

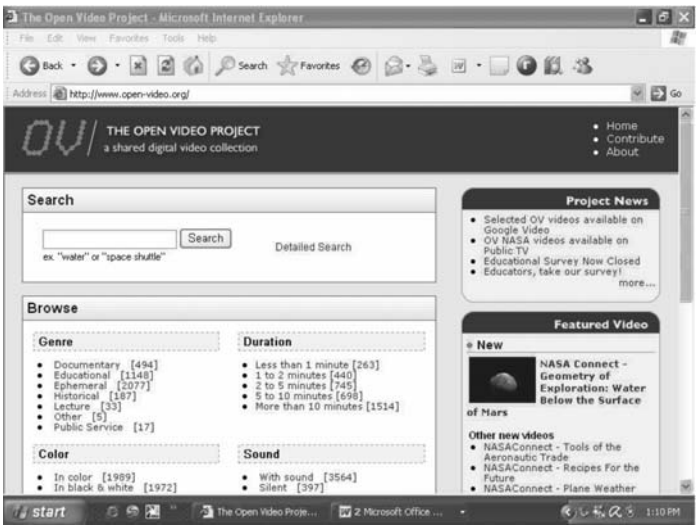
E-LIS is an electronic open access archive for scientific or technical documents, published or unpublished, in Librarianship, Information Science and Technology, and related application activities. E-LIS is an archive to deposit preprints, postprints and other LIS publications, finding and downloading documents in electronic format, offered as a free service to the international LIS community. The goal of the E-LIS archive is to promote communication in the field by the rapid dissemination of papers.

CogPrints: Cognitive Sciences E-print Archive is an electronic archive for self-archive papers in any area of Psychology, neuroscience, and Linguistics, and many areas of Computer Science (e.g., artificial intelligence, robotics, vision, learning, speech, neural networks), Philosophy (e.g., mind, language, knowledge, science, logic), Biology (e.g., ethnology, behavioral ecology, sociobiology, behaviour genetics, evolutionary theory), Medicine (e.g., Psychiatry, Neurology, human genetics, Imaging), Anthropology (e.g., primatology,



cognitive ethnology, archeology, paleontology), as well as any other portions of the physical, social and mathematical sciences that are pertinent to the study of cognition.

Open Video Project :



The Open Video Project is a shared digital video repository and test collection intended to meet the needs of researchers in a wide variety of areas related to digital video. The Open Video collection currently contains video or metadata for 1844 digitized video segments.

3. OAI PMH verbs

The OAI PMH has six request syntaxes that are used to send request to the data providers. These are:

- Identify
- ListMetadataFormats
- ListSets
- GetRecord
- ListIdentifiers
- ListRecords

The functions of these six verbs are given in the table below

Verb	Function
Identify	Retrieve information about repository
ListMetadataFormats	Retrieve the metadata format available from a repository
ListSets	Retrieve the set structure of a repository
GetRecords	Retrieves an individual metadata record from a repository
ListIdentifiers	Retrieves unique identifiers from a item
ListRecords	Harvest records from a repository

Table1. OAI PMH verbs with their functions

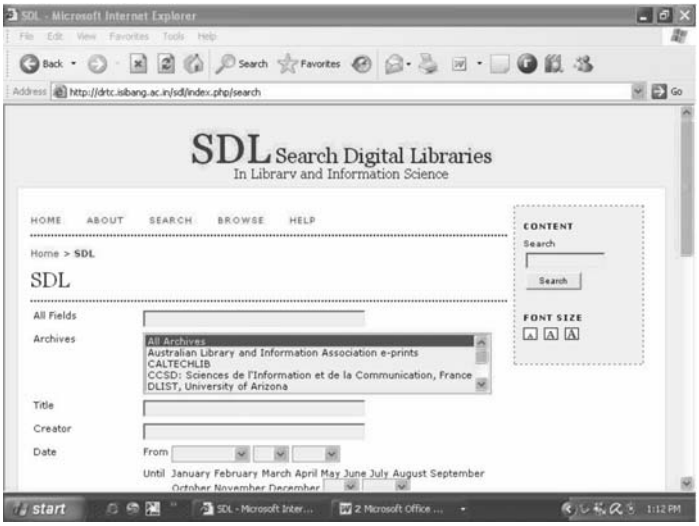
The service providers send the requests which are transmitted according to the rules of HTTP through the web. After receiving the request the data providers send the response in some valid XML format specified by the OAI PMH protocol. After receiving the response the service provider can understand who the specific metadata provider is, what the metadata format is. A digital repository can act as service provider or data provider or act as both.

4. Some Existing Metadata Harvester Services in India

In India also the efforts towards using and getting the services to the users by some institutional repository of some reputed institutions have been noticed. Some of the metadata harvester services of India are described below.

Search Digital Libraries (SDL):

It is a metadata service harvester launched by Documentation Research & Training Centre (DRTC) Bangalore. It is using PKP (Public Knowledge Project) Harvester software. It is concerned with selective harvesting to collect metadata records from the subject Library & Information Science (LIS) only. It is the second of its kind in the area of LIS in the world. Presently this service is harvesting more than 19,499 metadata records.



SJPI Cross Journal Search Service: The SJPI Cross Journal Search Service is a part of the SJPI project. The goal of this project is to improve the accessibility of scientific literature published in Indian journals by introducing an indexing system. A Sampling of journals is made OAI complaint using Open Journal System. A system (OJS). They are harvested by the PKP harvester in this site. This demonstrates a search service across multiple journals from a single point. This harvester indexes articles published in these journals and provides various ways of accessing them. Simple searching of keywords, author(s), title, abstract or index terms and advanced search are possible. The SJPI Harvester currently has 1047 papers from 13 journals indexed.

SEED (Search Engine for Engineering Digital-repositories):
The IIT Delhi has developed a number of disciple oriented Research Support Tools (RST), which accompanies individual research studies indexed from e-journal and conference paper websites covering a wide range of disciplines. The RST utilizes the study's metadata to search relevant open-access databases for related studies, theory, news, policies, and other resources, as well as offering access to the study's metadata and citation, to a personal portfolio. It uses PKP harvester. Simple and advanced search facility along with browsing capabilities is available. The Seed currently has 6176 papers from 4 archives indexed.

Open J-Gate:
Open J-Gate is an electronic gateway to global journal literature in open access domain. Launched in 2006, Open J-Gate is the contribution of Informatics (India) Ltd to promote OAI. Open J-Gate provides seamless access to millions of journal articles available online. Open J-Gate is also a database of journal literature, indexed from 3000+ open access journals, with links to full text at Publisher sites. It indexes articles from more than 3000 academic, research and industry journals, out of which more than 1500 are peer-reviewed scholarly journals. At present Open J-Gate indexes 3721 open access journals from different subject categories viz.Agricultural and Biological Sciences, Arts & Humanities, Basic Sciences, Biomedical Sciences, Engineering & Technology, Library & Information Sciences and Social & Management Sciences.

Knowledge Harvester of INSA:

Knowledge Harvester of INSA (Indian National Science Academy) is an experimental Open Access initiative. It indexes three archives viz. African Journal Online (currently has 248 journals), Archive of European Integration (currently has 5046 documents) and INSA Digital Library.

5. PKP: the Open Archives Harvesting software

One of the most widely used metadata harvester software is the PKP Open Archives harvester. PKP i.e. the Public Knowledge Project is dedicated to improving the scholarly and public quality of research. It operates through a partnership among the Faculty of Education at the University of British Columbia, the Simon Fraser University Library, the School of Education at Stanford University, and the Canadian Centre for Studies in Publishing at Simon Fraser University. Its research program is investigating the social, economic, and technical issues entailed in the use of online infrastructure and knowledge management strategies to improve both the scholarly quality and public accessibility and coherence of this body of knowledge in a sustainable and globally accessible form. It continues to be an active player in the open access movement, as it provides the leading open source software for journal and conference management and publishing. The PKP Open Archives Harvester is a free metadata indexing system developed by the Public Knowledge Project through its federally funded efforts to expand and improve access to research.

The PKP OAI Harvester allows us to create a searchable index of the metadata from Open Archives Initiative (OAI)-compliant archives, such as sites using Open Journal Systems (OJS) or Open Conference Systems (OCS). Public Knowledge Project has developed Open Journal Systems (OJS) which is a journal management and publishing system to expand and improve access to research. It has also developed Open Conference Systems (OCS) which is a free Web publishing tool that will create a complete Web presence for the scholarly conference. OCS will allow us to create a conference Web site, compose and send a call for papers, create a conference Web site, compose and send a call for papers, electronically accept paper and abstract submissions, allow paper submitters to edit their work etc.

Harvester version 2.x includes the following features:

- Ability to harvest OAI metadata in a variety of schemas (including unqualified DC, the PKP (Open Journal Systems/Open Conference Systems) Dublin Core extension, MODS, and MARCXML). Additional schema is supported via plugins.
- Flexible search interface that allows simple searching and advanced searching using cross walked fields from all harvested archives. Advanced searching of archives that share the same schema will be possible using fields as defined in the schema. When creating crosswalks for searching, administrator can define elements such as text, date, or HTML multiple select interface widgets.

- Ability to perform post-harvest and pre-indexing filtering/normalization on metadata.
- Searching is highly scalable (creates an inverted index for searching).

The PKP Open Archives Harvesting software is used by repositories like Search Digital Libraries (SDL), SEED of IIT Delhi, Digital Library of Information Science and Technology (DLIST), The University of Glasgow Open Archives Harvester, Petroleum Journals Online's Metadata Archive, UMS Repository, UIS Press Harvester etc.

6. Conclusion

To overcome the problem of lack of interoperability OAI protocol for Metadata Harvesting is an effective solution. After the release of version 2.0 of OAI-PMH is now possible to cover image, video, audio, and multimedia besides the various textual formats. Still there are a large number of archives which are not exposing their metadata to OAI-PMH for open access. Most of the e-journals which are using the OJS (Open Journal System) now are can be harvested by the OAI service providers. But at the same time OAI-PMH has not been able to provide the users the information of the non- OAI repositories which remain undiscovered. For the global proliferation of information we must have to adopt the OAI tools like OAI-PMH at the earliest.

References

1. Lagoze, Carl, Van de Sompel, Herbert, Nelson, Michael, and Warner, Simeon. *The Open Archives Initiative Protocol for Metadata Harvesting: Protocol Version 2.0 of 2002-06-14*. from http://www.openarchives.org/OAI/openarchivesprotocol.html
2. The Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 2.0 from http:/www.openarchives.org/OAI/openarchivesprotocol.html#item
3. Sompel, H. V. d., & Lagoze, C. (2000, February). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2). from http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html
4. Suleman, H., & Fox, E. A. (2001, December). A Framework for Building Open Digital Libraries. *D-Lib Magazine*, 7(12). from http://www.dlib.org/dlib/december01/suleman/12suleman.html
5. http://arxiv.org/
6. http://cogprints.ecs.soton.ac.uk/
7. http://drtc.isibang.ac.in/sdl
8. http://eprints.rclis.org/
9. http://eprint.iitd.ac.in/seed/viewarchive.php?id=6
10. http://metalis.cilea.it/
11. http://www.ncstrl.org/

13. http://www.caister.org/
14. http://www.openj-gate.com/
15. http://www.open-video.org/
16. http://pkp.sfu.ca/about/

ABOUT AUTHOR

Mr. Dipen Deka holds MLISc and has qualified UGC (NET)-JRF and presently undergoing research under the supervision of Prof. N Lahkar, Head of the DLIS, Gauhati University. He is also involved in teaching the students of MLISc course in the DLIS, Gauhati University. His areas of interest are digital libraries, OSS, digital preservation.