SEARCH ENGINE COVERAGE OF OPEN ACCESS CORPUS IN THE FIELD OF BIOTECHNOLOGY

*Rafiq Ahmad Rather
*Ms Rosy Jan
&
*Fayaz Ahmad Loan

ABSTRACT

The paper presents the results of an exploratory study conducted to find percentage of OAI-PMH compliant metadata harvested by two search engines (Google and Scirus) using two hundred and ten articles selected from DOAJ corpus. The first fifty results were analysed to gauge the presence of DOAJ articles in the field of Biotechnology with their rank and duplication among the retrieved set of results. The results reveal that Google harvest maximum articles (i.e. 76.67%) compared to Scirus while Scirus rank all the retrieved articles among the first ten results when Google fail by 6%.

KEYWORDS

OAI-PMH Compliant Resources, Metadata Harvesting, Open access resources, Biotechnology.

INTRODUCTION

The web is one of the youngest and fastest media growing exponentially. However, the major portion of it consisting scholarly publications pertain to journal archives, institutional repositories, databases, directories, digital libraries etc form part of the 'Deep web' i.e. not accessible at article or content level through traditional web search engines, though the archive or database may be itself accessible. This part of the web (deep web) contains valuable resources besides being 400 to 550 times larger than the commonly

*Rafiq Ahmad Rather, Rosy Jan & Fayaz Ahmad Loan (Research Scholars).The Department of Library and Information Science University of Kashmir, Srinagar, 190006, J&K, India.

defined www (web) (**Bright Corporation, 2006**). The components of the "deep web" represent significant institutional investment, yet their resources often remain hidden (**Sompel & Lagoze, 2000).**

A variety of techniques are developed for making the "deep web" accessible to enable researchers to find and access articles which would otherwise be unable to exploit them. One such frame work developed by Open Archives Initiative 'OAI' is Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH). The protocol is designed to enable transport of structured metadata from data repositories across the internet to build federated results (**Open Archives, 2006**). With the incoming of OAI-PMH, open access institutional repositories, databases, digital libraries etc are adopting the protocol to expose metadata about their resources to the scholarly world. At the same time search engines are becoming OAI-PMH compliant to enable them to index OAI-PMH resource corpus. However, search engines are not able to harvest exhaustively the resources of an OAI-PMH compliant repository or database. The present study demonstrate size of the metadata harvested by two popular search engines (Google and Scirus) from open access corpus selected from OAI-PMH Compliant directory (DOAJ) in the field of Biotechnology. The study may serve as an exploratory study to take up whole corpus of DOAJ for arriving at generic results.

LITERATURE REVIEW

A sizable literature describe features and metadata formats of recent concept of OAI-PMH but its successful implementation to facilitate interoperability between search engines and the deep web is meager. A study by **Sompel, Young and Hickey (2003)** describe innovative applications of OAI-PMH such as resource and metadata format and illustrate the usefulness of the OAI-PMH beyond the typical resource discovery using Dublin Core metadata. The author reveal that OAI-PMH provides a simple yet powerful framework for metadata harvesting and that OAI-PMH repositories have been directly overlaid with an interface that allows users to navigate the contained metadata by means of a web browser. **Hellgren (2004)** explores the implementation of the open archives initiative – metadata harvesting

protocol and the impact it may likely have on knowledge sharing .It reveals that users have come to expect instant and simple access to qualitative information resources through the use of Internet search engines. **Boston (2005)** present statistics on increased web usage focusing particularly on the collection of National Library of Australia. The author explores application of technologies such as the Open Archives Initiative Protocol for Metadata Harvesting to share deep web content through search engines and disclose that users can easily find information from the deep web using popular search engines. **Cole and Warner (2005)** provide an overview of emerging guidelines and best practices for OAI data providers. The authors present both protocol best practices and general recommendations for creating and disseminating high quality sharable metadata. The authors suggest that audience should be familiar with OAI-PMH having some experience with either data provider or source provider implementation. **Rodriguez, Bollen and Sompel (2005)** emphasize a deconstructed publication model in which the peer-review process is mediated by an OAI-PMH peer-review service using a social-network algorithm to determine potential reviewers. The authors advocate a set of peer-review specific metadata tags accompanying a pre-prints existing metadata records facilitating a unique repository that fits within the widely deployed OAI-PMH framework. **Xiang and Margan (2005)** describe the design and implementation of light weight protocols and open source tools including OAI-PMH. The authors describe how these protocols and tools are employed to collect, organize, archive and disseminate information freely available on the Internet. The seminal work by **McCown, Liu, Nelson and Zubair (2006)** evaluate three search engines namely Google, MSN and Yahoo for harvesting OAI-PMH resource corpus using 10 million records from 776 OAI-PMH repositories. The authors find that Yahoo index 65% followed by Google (44%) and MSN (7%) while as 21% of the resources are not indexed by any of the three search engines.

SCOPE

A plethora of search engines exist ranging from general to subject specific. However, the scope of the study is confined to two

search engines (Google and Scirus) both being OAI-PMH compliant. While Google belongs to General category, Scirus exclusively pertains to Science and Technology. The study is also limited to Directory of Open Access Journals (DOAJ) characterized by being OAI-PMH compliant, wide in scope and international in coverage. The study is further confined to the field of Biotechnology for English language only.

METHODOLOGY

The investigation was carried out in two stages. In the first stage search engines were selected and search articles drawn subsequently.

In the second stage search articles were run on the select search engines from 20th September to 10th October, 2006.

POPULATION SELECTION

DOAJ (2006) include 2401 Journals pertaining to different subjects. However, only 697 journals are searchable at article level and eight of them belong to Biotechnology comprising 265 articles. The fifty five articles in languages other than English were excluded from the study. Thus the population for investigation was reduced to 210 titles only.

TEST ENVIRONMENT

Each title was submitted to both the search engines in full words using simple mode of search. In case of a title that produces no results was reduced to its content bearing words and resubmitted. First fifty results were evaluated to determine the presence of the articles, their duplication and the rank provided by the respective search engine.

RESULTS AND DISCUSSION

The results analysed to determine the size of articles retrievable from DOAJ corpus (Table 1)

Table 1. Number of articles retrieved

N = 210

Search Engine	No. of Titles	
	Retrieved	***Not Retrieved***
Google	161 (76.67)	49 (23.33)
Scirus	80 (38.10)	130 (61.90)

*Figures in Parentheses indicate percentage

reveal that Google index 76.67% of the resources where as 23.33% are not retrievable through it. Scirus retrieve 38.10% but fail to retrieve 61.90% of the resources. Seventeen percent of the resources are not indexed by either of the search engines.

The results retrieved for their contents show that many titles harvested are repeated in results under different URLs in the first 50 hits (Table 2).

Table 2. Duplication in Results

N=210

Search Engine	Repeated Results
Google	39 (18.57)
Scirus	03 (1.43)

*Figures in Parentheses indicate percentage

It depicts that duplication is more prevalent in Google (18.57%) as compared to Scirus (1.43%).

The ranking status of the resources from the retrieved corpus indicate that Scirus rank 100% of the articles among the first 10 hits whereas Google show 94% among first 10 hits and 6% spill over ten hits (Table 3).

Table 3. Ranking of resources retrieved

Search Engine	Rank		Total
	Below 10	*Above 10*	
Google	151 (93.79)	10 (6.21)	161 (100)
Scirus	80 (100)	—	80 (100)

*Figures in Parentheses indicate percentage.

Recently published research show that Google index more web than other search engines like Yahoo, MSN, Altavista and Scirus. The most recent study (**McCown et al, 2006**) reveals Yahoo performing over Google in harvesting OAI-PMH compliant open access corpus. These studies however, do not remain valid for longer time due to dynamic nature of search engine algorithm and other policies. The present study reveals that Google adopt OAI-PMH to a great extent than Scirus. Google perform significantly better than Scirus in indexing OAI-PMH compliant DOAJ corpus. It indexes maximum of the DOAJ resources (76%) whereas Scirus couldn't harvest maximum of the resources (62%). Both the search engines perform well while ranking the resource. Scirus outperforms Google in ranking all the resources among the first ten hits. The limitation of Google to repeat results need to be addressed in its policy to come up to expectations of users particularly research scholars. The study need to be extended further to cover other facets of Science and Technology for well known search engines for understanding the impact of OAI-PMH protocol. The results may prove significant in improving upon the protocol besides usher in new tools and techniques for federating results. The study can further be extended to take up precision and recall of search engines in harvesting of metadata and ultimately may pave a way for a refined strategy for evolving digital library and search engine technology. A comparative study of impact of OAI-PMH protocol to the techniques and tools of digital libraries will open new vistas of R & D in future in the field of information science.

REFERENCES

Boston, Tony (2005). Exposing the Deep Web to increase access to library collections. Retrieved October 23, 2006 from http://ausweb.scu.edu.au/aw05/index

Bright Corporation (2006).The 'Deep Web': Surfacing Hidden Value. Retrieved October 14, 2006 from http://www.brightplanet.com/resources/details/deepweb.htm

Cole, Tim and Warner, Simeon M. (2005). OAI-PMH repositories: quality issues regarding metadata and protocol compliance. Retrieved October 25, 2006 from http://eprints.rclis.org/archive/00005502/

DOAJ (2006).Retrieved October 14, 2006 from http://www.doaj.org/

Hellgren, Timo (2004). OAI Compatibility: Exposing Metadata of Scientific Publications .Retrieved October 15, 2006 from http://www2.db.dk/NIOD/hellgren.pdf

Liu, Xiaoming (2002). A Scalable Architecture for Harvest –based Digital Libraries, D-Lib Magazine, 18 (11). Retrieved October 25, 2006 from www.dlib.org/dlib/november02/liu/11liu.html

McCown, F., Liu, Xiaoming, Nelson, M. L. and Zubair M (2006). Search Engine Coverage of the OAI-PMH Corpus. Internet Computing, 10 (2), 66- 73.Retrieved October 15, 2006 from library.lanl.gov/cgibin/getfile?LA-UR-05-9158.pdf

Open Archives (2006). Open archive initiative. Retrieved October 25, 2006 from http://www.openarchives.org

Rodriguez, Marko A, Bollen, Johan and Sompel, Herbert Van de (2004). The Convergence of Digital-Libraries and the Peer-Review Process. Retrieved October 20, 2006 from http://arxiv.org/pdf/cs.DL/0504084

Sompel, H. Van de and Lagoze, C. (2000). The Santa Convention of the Open Archives Initiative, D-Lib Magazine, 6(2).Retrieved October 22, 2006 from www.openarchives.org/documents/jcdl2001-oai.pdf

Sompel, Herbert Van de, Young, Jeffrey A and Hickey, Thomas B (2003). Using the OAI-PMH ... Differently, D-Lib Magazine, 9 (7/8).Retrieved October 25, 2006 from http://www.dlib.org/dlib/july03/young/07young.html

Xiang, X and Margan, E. L (2005). Leight- Weight Protocals and open source tools to implement Digital Library collections and services, D-Lib Magazine 11 (10). Retrieved October 17, 2006 from http://www.dlib.org/dlib/october05/morgan/10morgan.html