RETRIEVING MEDICAL LITERATURE from SELECT OAI-PMH COMPLIANT e-REPOSITORIES: A Case Study of Google and Yahoo

*Tariq Shafi

ABSTRACT

The study reports an exploratory investigation conducted on two search engines (Google and Yahoo) to find the retrieval percentage, duplication and rarking status of OAI-PMH Compliant resources harvested from five repositories in the field of Medical Sciences. The first twenty hits are analysed to reveal the findings in accordance with the laid down objectives for the study. The results show that Google is more comprehensive in retrieving OAI-PMH Compliant Medical Literature as compared to Yahoo. The study reveals that more duplication of results under different URLs in Google than in Yahoo. The results indicate that Google index 88.89% resources among the first 10 hits and 11.11% beyond the 10th hit, whereas Yahoo retrieves 82.61% results from the first 10 hits and 17.39% above the 10th hit.

KEYWORDS

OAI-PMH Compliant Resources, Metadata Harvesting, e-repositories, Medical Literature

INTRODUCTION

Searching on the internet today can be compared to dragging a net across the surface of the ocean. While a great deal may be caught in the net, there is still a wealth of information that is deep and therefore, missed. The reason being that most of the web's information is buried far down in the deep web - the part of the

Librarian, Kashmir Paradise College of Education. Nishat, Srinagar e-mail: tariq_lis@yahoo.co.in

web that is typically hidden from web search engines. This part of the web is a vast reservoir of internet content that is 500 times larger than the surface web (**Bright Corporation, 2006**). The components of the "deep web" represent significant institutional investment, yet their resources often remain hidden **(Sompel & Lagoze, 2000).**

The inability of the web search engines to crawl the deep web which approximately constitutes 70% of the existing web, has given rise to many techniques, 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). It is a protocol (OAI-PMH) used to harvest (or collect) the metadata descriptions of the records in an archive so that services can be built using metadata from many archives (**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.

LITERATURE REVIEW

The concept of OAI-PMH emerged during the late 1990s. Though the movement is quite recent, the literature describing successful implementation to facilitate interoperability between search engines and the deep web is quite sizeable. 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 protocol best practices and general recommendations for creating and disseminating high quality sharable metadata. They 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 most recent study **(McCown et al, 2006)** reveals Yahoo performing over Google in harvesting OAI-PMH compliant open access corpus. 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. Study by **(Markland, 2006)** highlights the efficiency of Google and Google Scholar in retrieving the data from 26 U.K Institutional Repositories, covering a wide range of subject areas. The study reveals that full title search and sophisticated harvesting services could prove a better option for the scholars.

SCOPE

The scope of the Study is limited to five OAI-PMH Compliant Medical Repositories having English Language databases, for assessing coverage of the corpus in two Search Engines viz. Google and Yahoo.

Objectives

The following objectives are laid down for the study:

a) To measure Search Engine coverage (retrieval percentage) of the OAI-PMH corpus in the field of Medical Science.

b) To determine the Repetition of results under different URL's.

c) To determine Rank of the resources retrieved from the select repositories.

METHODOLOGY

The study was carried out in the following three stages:

I. Selection of OAI-PMH Compliant Medical Repositories

The following five OAI-PMH Compliant Medical Repositories having English Language database were selected by Purposive or Judgment sampling from the four Registries[1] of OAI-PMH Compliant Repositories:

	Name of Repository	Base URL
1.	BioMed Central	http://www.biomedcentral.com/oai/2.0
2.	DSpace at the University of Washington Health Sciences Libraries	http://dspace.hsl.washington.edu/dspace-oai

[1] (a) The open archives list of registered OAI-PMH repositories (www.openarchives.org/register/browsesites) (b)The OAI registry at the University of Illinois at Urbana-Champaign (http://gita.grainger.uiuc.edu/registry/info.asp) (c)The Celestial OAI registry (http://celestial.eprints.org) (d) Eprint's Institutional Archives Registry (http://archives.eprint.org)

3.	MeDSpace at Duke University Medical Center Library and Archives	http://dspace.mclibrary.duke.edu /dspace-oai
4.	OpenMED@NIC	http://openmed.nic.in/perl/oai2
5.	Washington University School of Medicine	http://linpub1.wustl.edu/dspace-oai

II. Harvesting of the Resource Corpus from select OAI-PMH Compliant Medical Repositories

Here 10% of the resource corpus was harvested from each of the select repositories by Quasi-Random Sampling.

III. Checking the harvested resource corpus (titles) with select Search Engines

In this stage, the harvested resource corpus was run on the select search engines. First twenty hits were evaluated to gauge the presence of titles. Since Search engine coverage refers to the comprehensiveness of a search engine to index the contents of the web, therefore while analysing retrieval percentage of select Search Engines, the study took into consideration their total retrieved results. And to assess search engine coverage of the OAI-PMH Corpus in the field of Medical Science, 1310 titles were harvested from select OAI-PMH Compliant Medical Repositories. Among these, 10% of the titles i.e. 131 titles were selected by Quasi random Sampling and then Google and Yahoo were queried to see if they had indexed the select titles. The data thus collected was compiled, analysed and presented in tabular and graphic form to reveal the findings in accordance with the laid down objectives of the study.

DISCUSSION

Search Engine Coverage

Among 131 select titles, Google retrieved 108 titles (82.44%) and Yahoo 92 titles (70.23%). Google missed 23 titles (17.56%), and Yahoo 39 titles (29.77%).Table 1 depicts that Google and Yahoo performance in retrieving titles from 4 repositories viz, *OpenMED@NIC, MeDSpace at Duke University Medical*

Center Library and Archives, DSpace at the University of Washington Health Sciences Libraries, and *Washington University School of Medicine* with respective retrieval percentage of 89.47%, 89.13%, 82.61% and 60.00%. Yahoo outperformed Google in BioMed Central only (78.79% vs. 75.76%). The highest non-retrieval percentage is observed in *Washington University School of Medicine* (Google 40% and Yahoo 50%), while OpenMed with least non-retrieval percentage (Google 10.53% and Yahoo 15.79%). .

Table 1: Search Engine Coverage (Retrieval Percentage)

S.No.	Repository	Select Titles	Google		Yahoo	
			R	NR	R	NR
1	BioMed Central	33	25 (75.76)	8 (24.24)	26 (78.79)	7 (21.21)
2	DSpace at the University of Washington Health Sciences Libraries	23	19 (82.61)	4 (17.39)	16 (69.57)	7 (30.43)
3	MeDSpace at Duke University Medical Center Library and Archives	46	41 (89.13)	5 (10.87)	29 (63.04)	17 (36.96)
4	OpenMED@NIC	19	17 (89.47)	2 (10.53)	16 (84.21)	3 (15.79)
5	Washington University School of Medicine	10	6 (60.00)	4 (40.00)	5 (50.00)	5 (50.00)
TOTAL		**131**	**108 (82.44)**	**23 (17.56)**	**92 (70.23)**	**39 (29.77)**

* *Figures in parenthesis indicate percentage*

Composite Retrieval

To determine the composite retrieval percentage of the select search engines, those titles were taken into account which were either retrieved by both of the search engines or were not retrieved by either of the engines.

While analysing the retrieved results of select search engines, it is observed that 82 titles (62.66%) were retrieved by both the engines.The composite retrieval of select search engines is highest in OpenMED@NIC (84.22%) followed by BioMed Central (66.67%), while the lowest percentage (50.00%) is found in Washington University School of Medicine.(Table2.)

The select search engines are not prolific in retrieving 13 titles (9.92%) among the select titles (131) for the study. It is obvious from the study that select search engines perform poor in retrieving titles from Washington University School of Medicine (40.00%), followed by BioMed Central (12.12%) and OpenMED@NIC (10.53%) respectively.

Table 2: Composite Retrieval Percentage

S.No.	Repository	Select Titles	Composite Retrieval Percentage	
			Retrieved by Both	Retrieved by None
1	BioMed Central	33	22 (66.67)	4 (12.12)
2	DSpace at the University of Washington Health Sciences Libraries	23	13 (56.52)	1 (4.35)
3	MeDSpace at Duke University Medical Center Library and Archives	46	26 (56.52)	2 (4.35)
4	OpenMED@NIC	19	16 (84.22)	2 (10.53)
5	Washington University School of Medicine	10	5 (50.00)	4 (40.00)
	TOTAL	**131**	**82 (62.60)**	**13 (9.92)**

Duplication (Repeated Results)

Many results are repeated under different URLs in the first 20 search hits of the search engines. Those hits were reviewed for their contents and analysed for all select titles.

While comparing the duplicated results of Google and Yahoo, a duplication of 268 hits (12.41%) among 2160 evaluated results for 108 retrieved titles is found in Google. Yahoo results in 126 duplicate hits (6.85%) occurring in 1840 evaluated results for 92 retrieved titles. (Table 3).

Table 3: Duplicate results of Google & Yahoo

N = 131

Search Engine	Titles Retrieved	Hits Evaluated	Repeated Hits
Google	108	2160	268 (12.41)
Yahoo	92	1840	126 (6.85)

* *Figures in parenthesis indicate percentage*

Ranking

The ranking status of the retrieved results indicate that Google rank 88.89% of the results among the first 10 hits, while as 11.11% spill over the 10th hit. On the other hand, Yahoo grade 82.61% results between first 10 hits, whereas 17.39% go beyond the 10th hit. (Table4).

Table4: Ranking of Resources Retrieved

Search Engine	Retrieved Results	Rank	
		Below 10	Above 10
Google	108	96 (88.89)	12(11.11)
Yahoo	92	76 (82.61)	16(17.39)

* *Figures in parenthesis indicate percentage*

CONCLUSION

The Open Archives Initiative Protocol for Metadata Harvesting has emerged as a practical foundation for digital library interoperability. It can be used by a variety of communities who are engaged in publishing content on the web, as the OAI protocol has great potential to expose hidden resources via the web. The open Archives Metadata Harvesting Protocol offers a new bridge to bring innovation in networked information services and applications out of the research community more rapidly than has been the case in the past. The OAI Protocol for Metadata Harvesting offers the prospect of resource discovery tools far beyond what is currently available to users of the web via standard Search Engines, none of which (currently) make any substantial use of metadata. This paper illustrates how existing information about available resources can be repurposed fairly, easily and cheaply using standard tools. The study has also highlighted the limitation of Google to repeat results that need to be addressed in its policy to come up to expectations of users particularly research scholars. This paper has also identified issues that need to be addressed in order to deploy a truly interoperable framework for resource harvesting based on the use of OAI-PMH addressing scenarios in which large resources are to be harvested and conveying rights pertaining to harvestable resources.

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 January 15, 2007, 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/

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

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 January 18, 2007 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., & Lagoze, C. (2000). The Santa Convention of the Open Archives Initiative, D-Lib Magazine, 6(2).Retrieved January 22, 2007, 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). Light- Weight Protocols 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