The International JOURNAL of KNOWLEDGE, CULTURE & CHANGE MANAGEMENT

Volume 8

An Architecture for Efficient Resource Discovery with Metadata Harvesting in a Multidisciplinary Distributed Repository

Jayan C Kurian, M.G. Sreekumar, Dion Hoe-Lian Goh, Diljit Singh, Abrizah Abdullah and Joy Lynn Wheeler

COMMON GROUND

www.management-journal.com

THE INTERNATIONAL JOURNAL OF KNOWLEDGE, CULTURE AND CHANGE MANAGEMENT
http://www.Management-Journal.com

First published in 2008 in Melbourne, Australia by Common Ground Publishing Pty Ltd
www.CommonGroundPublishing.com.

ISSN: 1447-9524
Publisher Site: http://www.Management-Journal.com

THE INTERNATIONAL JOURNAL OF KNOWLEDGE, CULTURE AND CHANGE MANAGEMENT is a peer refereed journal. Full papers submitted for publication are refereed by Associate Editors through anonymous referee processes.

Typeset in Common Ground Markup Language using CGCreator multichannel typesetting system
http://www.CommonGroundSoftware.com.

An Architecture for Efficient Resource Discovery with Metadata Harvesting in a Multidisciplinary Distributed Repository

Jayan C Kurian, Nanyang Technological University, Singapore, SINGAPORE
M.G. Sreekumar, University of Malaya, Kuala Lumpur, MALAYSIA
Dion Hoe-Lian Goh, Nanyang Technological University, Singapore, SINGAPORE
Diljit Singh, University of Malaya, Kulala Lumpur, MALAYSIA
Abrizah Abdullah, University of Malaya, Malaysia, Kuala Lumpur, MALAYSIA
Joy Lynn Wheeler, Nanyang Technological University, SINGAPORE

Abstract: The profusion of non-relevant information for a given query on the Web explains the pressing need for formulating ebullient strategies for pertinent Web resource discovery and retrieval. One of the major requirements for effective document retrieval is its diligently encoded metadata. At the same time metadata standards to be followed for annotating documents from large collections are pretty complex. This is because the standardized global metadata cannot represent all the elusive forms of document metadata for improved retrieval ranking. In this context, we propose an approach to facilitate document retrieval from multidisciplinary domains where each belonging to discrete domains would be indexed in a segregated instance of a repository. This would facilitate document metadata customization for each specific discipline by adding specific metadata themes. Since the approach retains the standard metadata schema in addition to the customized metadata schema, it would result in enhanced resource discovery. The metadata retrieval process will be supported by an extended protocol for metadata harvesting (X-PMH) [1] and will be implemented in each repository. The extended metadata harvesting approach has been used to tie together the metadata customization components made at various repository instances. The proposed framework could be integrated into Open Digital Libraries (ODLs) [2] and shall serve as an intrinsic model that adds value in the context of multidisciplinary metadata simplicity, maintenance, and descriptive metadata availability in the event of repository instance failures. Our approach is to implement this cost-effective architecture using the PKP-OAI (Public Knowledge Project – Open Archive Initiative) [3,4] harvester on DSpace [5], an open source digital repository platform that supports metadata harvesting in its innate form. Once this is fully achieved, a federated search build upon such repository instances using open source technologies [6] would yield promising results in the context of information retrieval.

Keywords: Metadata Harvesting, Federated Searching, Information Retrieval, Multidisciplinary Distributed Repository, Public Knowledge Project, Open Archives Initiative, Open Digital Libraries

Introduction

THE ABUNDANCE OF heterogeneous information among multidisciplinary organizations has made the pervasive task of managing information a complex process. Hence in an information management context that involves aggregating relevant information and disseminating to the right audience, repositories play a magnanimous role due to their open-source nature. Thus information management and repositories are highly interrelated in a library context and cannot be seen as segregated and indolent in nature. To support this cohesive attainment towards scholarly information processing and dissemination, institutional repositories play a significant role. Institutional repositories, being developed as a major source of scholarly dissemination based on open-access policies, have become an integral component of almost all libraries especially in a multidisciplinary educational environment. We refer to a multidisciplinary repository (Open Access Archive) as a collection of peer-reviewed post-prints of scholarly articles, image and video objects, learning objects, student theses, and industrial attachment reports. In fact, the present digital libraries depend on repository collections to garner, process, and disseminate scholarly information, which otherwise would not have achieved its full potential in terms of scholarly dissemination. The crux of any institutional repository includes support for the dissemination of materials that include research papers, technical articles, theses, and working papers. In addition to that, the basic principles that govern the outreach of institutional repositories include support for learning objects, corporate assets, and granular content [7]. Learning objects includes organizational manifested materials (e.g. lecture notes), corporate assets include organizational records (e.g. annual reports), and granular content describes organizational content to its minute details

THE INTERNATIONAL JOURNAL OF KNOWLEDGE, CULTURE AND CHANGE MANAGEMENT,
VOLUME 8, 2008
http://www.Management-Journal.com, ISSN 1447-9524

(e.g. file level metadata description rather than compound object level description). Thus repositories play a significant role by supporting the scholarly communication lifecycle model of Roosendaal and Guertz that includes registration, certification, awareness, archiving, and rewarding [7].

The popular open source repository packages include Archimede, CDSware, DSpace, EPrints.org, Fedora, and OPUS. Archimede [8] was designed for the preservation and dissemination of scholarly content of university level research communities where as CDSware [9] initially started with the high energy physics research document collection and later traversed into a collection of general library resources. EPrints [10] also was based on the concept of open-policy dissemination of research content and the Fedora repository [11] brought in the concept of a scalable digital library system with an object model supporting multiple representations of unique digital object [7]. OPUS [12] was developed keeping in mind the concept of an information system for publishing staff and student electronic documents. DSpace [5] evolved as a collaborated project between Hewlett Packard Labs and MIT Libraries and has a wide community base with several volunteered committers for its evolving development. We base our further study and proposal of a multidisciplinary distributed architecture on this widely accepted open-source software due to our familiarity with the software and its implementation in our affiliate institutes.

Two of the major value adding features of OA archives, among many others, are their Internet presence (omnipresence) and interoperability. The interoperability feature keeps all OA archives virtually a single digital library system wherein they share their metadata through some common services called metadata harvesters or service providers using the OAI-PMH protocol. A number of tools are now available for starting such services and the PKP archive harvester is the appreciated and simple one used by many.

Research Issue

There is ample evidence of increasing acceptance of institutional repositories world wide. For instance, the Directory of Open Access Repositories (Open DOAR) maintained by the University of Nottingham, UK has recently crossed the 1100 listings [13]. At the same time the growth in the number of repositories, their distributed nature, size and diversity also make the quality of metadata an increasingly important issue [14]. Bruce and Hillmann acknowledge the difficulties in defining metadata quality [15]. They have identified seven metadata quality criteria: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. And it requires both human expertise and appropriate tool support for accessing these criteria in diverse collections [16,17]. Metadata quality is of paramount importance to a repository and the metadata completeness is a mark of the record quality of the repository.

In a multidisciplinary repository context effective representation of digital objects would be facilitated by various metadata schemas. This leads to a lack of universal representation schema that supports diverse digital objects. In this context we propose a harvesting-cum-indexing approach supported by the 2.0 version of the protocol for metadata harvesting (OAI-PMH 2.0) [18] and facilitated by the PKP-OAI (Public Knowledge Project – Open Archives Initiative) harvester [19]. The advantage of using this approach is extensible metadata schema and that can be used to represent diverse objects in a multidisciplinary repository. This model adds value in the context of multidisciplinary metadata simplicity, maintenance, and descriptive metadata availability in the event of repository instance failures. This also relieves from defining a single metadata schema encompassing the requirements for effectively describing variety of resources as well as large-scale resource intensiveness of central repositories. An additional research issue would to provide an interface by which a user can search harvested metadata based on encoded schema representation (e.g. Dublin core, Learning Object Model etc). This would stress on adaptive retrieval of digital resources

Proposed Case Study

The two affiliate institutes in this project are University of Malaya, Malaysia and the Nanyang Technological University, Singapore. Each university has built its own individual repository based on specific collection themes. Metadata schema may be tuned according to individual collection requirement. Harvesting can be applied between local repositories that act as data providers. Each university can also act as a service provider of its local collections. Since both universities can act as service providers metadata harvesting can be done between individual repositories. The proposed architecture is shown in figure 1.

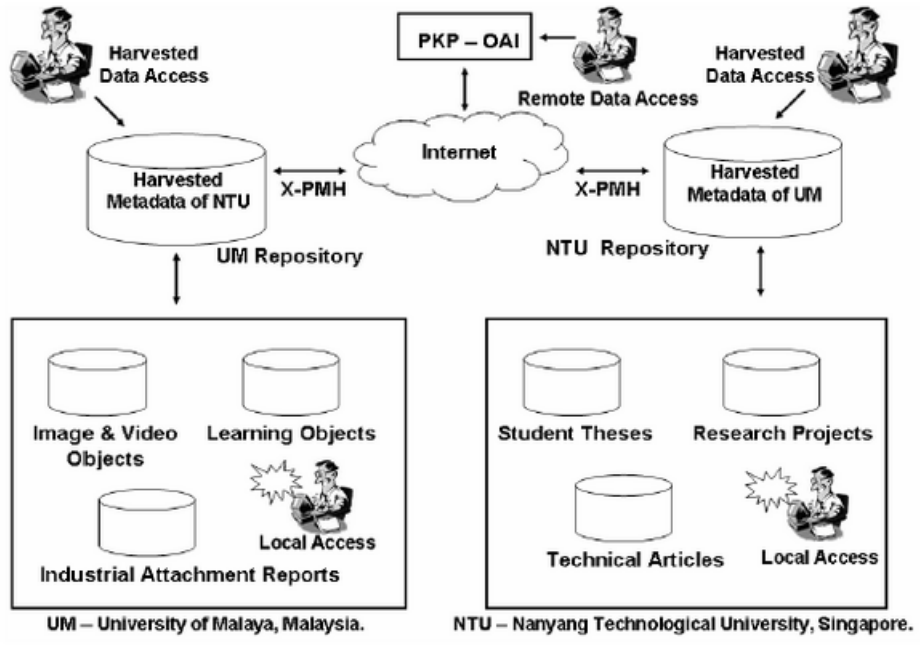

Figure 1

As an initial step we would be investigating how different metadata schemas are interoperable by performing a migration between distributed repositories. To perform this we utilize a repository script implemented in the DSpace software and schedule the script that would automatically export digital objects from one instance in a package format to be easily imported into another instance. Before importing into a second instance some amount of data cleansing has to be done that would take into account the schema representation of the second instance. At a later stage this may be automated to represent complex objects (consisting of more than one object) in different schema representations in different local repositories. The sample scheduler script that would be used in conjunction with the main DSpace Export feature is given below.

```
set DayDateMonthYear=%date:~0,3%%date:~7,2%-%date:~4,2%-%date:~10,4%
set TimeinHours=%time:~0,2%
set TimeinMinutes=%time:~3,2%
set FolderNamewithDayTime=%DayDateMonthYear%-%TimeinHours%hr%TimeinMinutes%mts
md E:\%FolderNamewithDayTime%
cd E:\dspace\bin
dsrun org.dspace.app.itemexport.ItemExport -t
COLLECTION -i 123456789/135 -d E:\%FolderNamewithDayTime% -n 1
```

Once this process is over, the open source PKP archive software will be installed, configured and customized to harvest the metadata records from both the UM and NTU repositories. The PKP harvester shall act as a common index having search and browse facilities over the Internet. The index service will be tested for its functionality and correctness over a period of time and subsequently more repositories will be added to the index.

Conclusion

In this on-going research work that is in its preliminary stage we have been successful in scheduling semi-automatic migration between local repository instances. Further work needs to be done in investigating metadata interoperability between repository instances. Another area to be investigated is metadata harvesting among local repositories to act as a data and service provider among affiliate institutes. We are hopeful that the PKP archives harvester shall be able to meet this efficiently and effectively. Once this model evolves as a promising initiative we plan to extend this with other institutes having similar interest in open source movement.

References

[1] Suleman, H., Fox, E.A. “Leveraging OAI Harvesting to Disseminate Theses”. Library Hi Tech, 21(2), pp. 219-227, 2003.
[2] Fox, E.A., Suleman, H., Luo, M. “Building digital libraries made easy: toward open digital libraries”, In Proceedings of the 5th International Conference of Asian Digital Libraries, 2002.
[3] The Public Knowledge Project, accessible at: http://pkp.sfu.ca/
[4] Suleman, H and Fox, E.A. “Designing protocols in support of digital library componentization”. In Proceedings of the European Conference on Digital Libraries, pages 568–582, Rome, Italy, 2002
[5] Dspace, accessible at: http://www.dspace.org/
[6] Library Find Project, accessible at: http://www.libraryfind.org/
[7] Richard Jones, Theo Andrew, John MacColl, “The Institutional Repository”, Chandos Publishing, 2006
[8] Archimede, available at: http://www.bibl.ulaval.ca/archimede/index.en.html
[9] CDSware, available at: http://cdsweb.cern.ch/
[10] EPrints.org, available at: http://www.eprints.org/
[11] Fedora, available at: http://www.fedora.info/
[12] OPUS, available at: http://elib.uni-stuttgart.de/opus/index.php?la=en
[13] OpenDOAR available at http://www.opendoar.org
[14] Nicholas, David M. et al. “A tool for metadata analysis”. Working Paper 2/2008, Department of Computer Science, University of Waikato. 2008.
[15] Bruce, T.R. and Hillmann, D.I. “The continuum of metadata quality: defining, expressing, exploiting”. In: Metadata in Practice, American Library Association, Chicago, IL. 238-256. 2004.
[16] Beall, J. “Metadata and data quality problems in the digital library”. Journal of Digital Information. 6(3) 2005.
[17] Ochoa, X and Duval, E. “Towards automatic evaluation of learning object metadata quality”. In: Advances in Conceptual Modeling – Theory and Practice, ER 2006 Workshops BP-UML, CoMoGIS, COSS, ECDM, OIS, QoIS, SemWAT, Springer, 372-381. 2006.
[18] OAI-PMH 2.0, available at http://www.openarchives.org/OAI/openarchivesprotocol.html
[19] PKP Harvester 2.x, available at http://pkp.sfu.ca/?q=harvester

About the Authors

Jayan C Kurian
Nanyang Technological University, SINGAPORE

Dr. M.G. Sreekumar
University of Malaya, MALAYSIA

Dr. Dion Hoe-Lian Goh
Nanyang Technological University, SINGAPORE

Dr. Diljit Singh
University of Malaya, MALAYSIA

Dr. Abrizah Abdullah
University of Malaya, Malaysia, MALAYSIA

Joy Lynn Wheeler
Joy Wheeler has worked at the NTU Library in the Library Technology Systems Division for the past 2 years. She splits her time as a subject librarian for the Art Design & Media Library.

The International JOURNAL of KNOWLEDGE, CULTURE & CHANGE MANAGEMENT

COMMON GROUND

Please visit the Journal website at http://www.Management-Journal.com
for further information about the Journal or to subscribe.