Implementation of Metadata Harvesting of Scientific and Scholarly Research Journal's Content

Pavel Simek[1], Jan Jarolimek, Jiri Vanek, Michal Stoces[2]

[1]Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic, e-mail: simek@pef.czu.cz

Abstract. In our paper we discuss basic objectives, methods and results of the implementation of metadata harvesting from the journal Agris on-line. The international journal Agris on-line Papers in Economics and Informatics is a scholarly open access, blind peer-reviewed, interdisciplinary, and fully reviewed scientific journal, which is published quarterly by the Faculty of Economics and Management at the Czech University of Life Sciences Prague (CULS). The metadata format used is the Dublin Core metadata standard. For metadata harvesting from the Agris on-line repository we use the Open Archives Initiative Protocol for Metadata Harvesting standard, version 2.0. Creators and administrators of the Agris on-line Papers in Economics and Informatics used their own solution for metadata storage and services for the metadata harvesting. The solution is opened to other repositories of the CULS.

Keywords: repository, open archive, metadata, OAI-PMH, Dublin Core, open access

1 Introduction

The number of publications made freely accessible on the Web by institutions joined in Open Archive Initiatives such as libraries, research institutions, scientific and cultural archives, has been constantly growing (Bellini, 2010). Libraries have been trying to find a faster and better way to provide access to resources they hold or resources available elsewhere (Han, 2011). The different content of different repositories can be described via metadata. Metadata is data on data. Metadata can be used for almost any content. Texts, pictures, music, movies, web pages, software (as content), etc. – all these materials can use metadata.

In July 1999, Paul Ginsparg, Rick Luce and Herbert Van de Sompel sent out a Call for Participation (Ginsparg, 1999) in a meeting exploring cooperation among scholarly e-print archives. The meeting, held in October 1999 in Santa Fe, and originally called the Universal Preprint Service meeting, led to the establishment of the Open Archives initiative (OAI) (Ginsparg, 1999). The goal of the OAi is to contribute to the transformation of scholarly communication in a concrete way. The proposed vehicle for this transformation is the definition of technical and supporting

In: M. Salampasis, A. Matopoulos (eds.): Proceedings of the International Conference on Information and Communication Technologies

for Sustainable Agri-production and Environment (HAICTA 2011), Skiathos, 8-11 September, 2011.

organizational aspects of the open scholarly publication framework on which both free and commercial layers can be established (Van de Sompel, 2000).

Some institutions build central metadata repositories (archives). Those larger repositories are filled from other smaller archives. The smaller archives and the central larger repository must provide some tools (applications) for automatic metadata harvesting. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has found widespread adoption for exchanging metadata from agrarian sector (Haslhofer, 2010). It is 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. OAI-PMH is based on a client–server architecture, in which harvesters request information on updated records from repositories.

Commercial search engines have started using OAI-PMH to acquire more resources, too. Google is using OAI-PMH to harvest information from the National Library of Australia Digital Object Repository. In 2004, Yahoo! acquired content from OAIster (University of Michigan) that was obtained through metadata harvesting with OAI-PMH. NASA's Mercury: Metadata Search System uses OAI-PMH to index thousands of metadata records from Global Change Master Directory (GCMD) everyday (McCown, 2006).

2 Data and methods

The international journal AGRIS on-line Papers in Economics and Informatics is a scholarly open access, blind peer-reviewed, interdisciplinary, and fully reviewed scientific journal. The journal is published quarterly by the Faculty of Economics and Management at the Czech University of Life Sciences Prague (CULS). AGRIS on-line Papers in Economics and Informatics covers all areas of agriculture and rural development: agricultural economics, management, agribusiness, agrarian policy, information and communication technologies, information systems, e-business, social economy and rural sociology. The journal provides a leading forum for an interaction and research on the above-mentioned topics of interest. The journal serves as a valuable resource for academics, policy makers and managers seeking up-to-date research on all areas of the subject.

The papers, which are to be published in Agris on-line Papers of Economics and Informatics, will be revised by the executive editor and the chief editor before being sent to two reviewers. The reviewers propose one of four recommendations:

1. accepted without revision
2. minor revision + comments
3. major revision + comments
4. reject + comments

Revised papers are checked by the executive again (to check whether the authors had revised their papers). All the accepted papers are checked by a member of the Editorial Board before publishing. In total they are reviewed by 5 people before being published:

1. executive editor
2. chief editor
3. the first reviewer
4. the second reviewer
5. editorial board member

There are more than sixty scholarly papers in pdf format in the journal repository and papers are - as mentioned above - inserted to the repository quarterly. The journal is available at http://online.agris.cz.

There are two requirements for the successful integration of the content of the repository.

1. The content of the journal repository must be described via metadata, because only metadata can present the basic content of each paper relevantly.
2. The journal repository must support the system for metadata harvesting.

A typical metadata harvesting approach using OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting) provides an application-independent interoperability framework based on metadata harvesting (Van de Sompel, 2004). There are two classes of participants in the OAI-PMH framework:

1. Data Providers administer systems that support the OAI-PMH as a means of exposing metadata; and
2. Service Providers use metadata harvested via the OAI-PMH as a basis for building value-added services.

A resource is the object or "stuff" that metadata is "about". The nature of the resource, whether it is physical or digital or stored in the repository or is a constituent of another database, is outside the scope of the OAI-PMH. An item is a constituent of a repository from which metadata about a resource can be disseminated. The metadata may be disseminated on the fly from the associated resource, transferred from any canonical form, actually stored in the repository, etc. A record is metadata in a specific metadata format. The record is returned as an XML-encoded byte stream in the response to the protocol request to disseminate a specific metadata format from a constituent item (Open Archive Initiative, 2008).

A harvester is a client application that issues OAI-PMH requests. The harvester is operated by a Service Provider as a means of collecting metadata from the journal repository. The journal repository is a network accessible server that can process the 6 OAI-PMH requests and the journal repository is managed by a Data Provider to expose metadata to harvesters. The operator of the Agris on-line Papers in Economics and Informatics repository is within Data Provider role.

The OAI-PMH solution will be opened and available to other repositories of the CULS with scholarly content.

3 Results

The journal repository uses for metadata basic Dublin Core standard which includes 15 recommended elements (Dublic Core Metadata Initiative, 2010).

However the repository does not use all fifteen elements but thirteen only. The list of all the used elements is as follows:

1. Coverage
2. Creator
3. Date (YYYY-MM-DD)
4. Descriptor
5. Format
6. Identifier
7. Language
8. Publisher
9. Right
10. Source
11. Subject
12. Title
13. Type

Creators and administrators of the Agris on-line Papers in Economics and Informatics used their own solution for metadata storage and service for metadata harvesting. The solution provides metadata for all the OAI-PMH requests:

1. Identify – this verb is used to retrieve information about the Agris on-line repository.
2. ListMetadataformats – this verb is used to retrieve the metadata formats available from the Agris on-line repository.
3. ListSets – this verb is used to retrieve the set structure of the Agris on-line repository.
4. ListIdentifiers – this verb is an abbreviated form of ListRecords, retrieving only the headers of the records.
5. ListRecords – This verb is used to harvest records from the Agris on-line repository.
6. GetRecord – this verb is used to retrieve an individual metadata record from the Agris on-line repository.

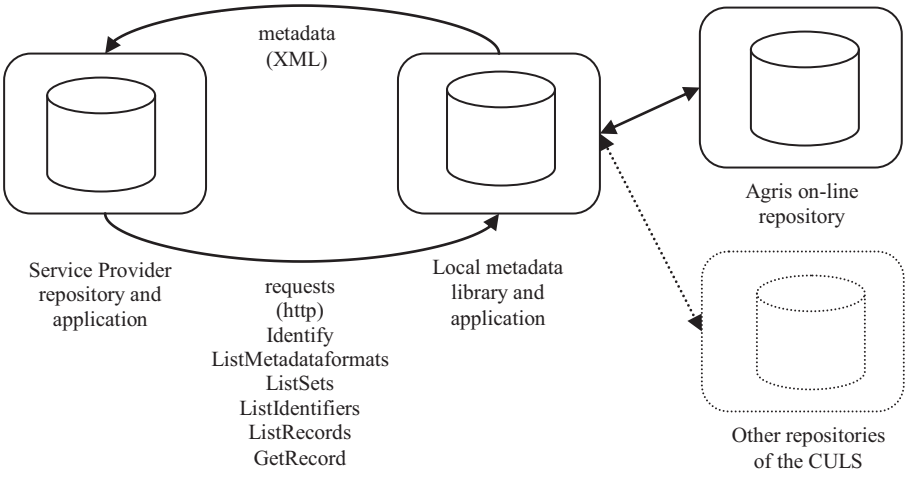

Fig. 1. OAI-PMH implementation at the Faculty of Economics and Management, the Czech University of Life Sciences Prague.

During the analysis of a possible OAI-PMH implementation it was found that it is not possible to use an open source solution such as DSpace or Drupal. DSpace is a perfect software of choice for academic, non-profit, and commercial organizations building open digital repositories. It is free, easy to install and completely customizable to fit the needs of any organization. But the technical environment at the Czech University of Life Sciences Prague does not support the technical platform DSpace Software needs, especially databases. This is the main reason why we created our own solution for metadata harvesting.

Creators and administrators of Agris on-line wanted to prepare a general solution, which would be usable for other repositories of the CULS. They have prepared a local metadata library where metadata from Agris on-line are archived. There are services for synchronizing of metadata between the local metadata library and Agris on-line repository or another university repository. The application of local metadata library supports the OAI-PMH version 2.0 as a means of exposing metadata.

Archived metadata are classified into sets. The specification of the set for the electronic scientific journal is oai:aol and its name is Agris on-line Papers in Economics and Informatics.

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2011-04-20T10:55:46Z</responseDate>
  <request verb="GetRecord" identifier="oai:oai.agris.cz/?id=2"
         metadataPrefix="oai_dc">http://oai.agris.cz</request>
  <GetRecord>
   <record>
    <header>
     <identifier>oai:oai.agris.cz/2</identifier>
     <datestamp>2009-09-30T00:00:00Z </datestamp>
     <setSpec>oai:aol</setSpec>
    </header>
    <metadata>
     <oai_dc:dc
       xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
       xmlns:dc="http://purl.org/dc/elements/1.1/"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
       http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:title>Cash flow as an important solvency determinant of agricultural enterprises in
       the Slovak Republic</dc:title>
      <dc:creator>Gurčík, L.</dc:creator>
      <dc:creator>Miklovičová, J.</dc:creator>
      <dc:creator>Miklovičová, S.</dc:creator>
      <dc:subject>Agricultural enterprises</dc:subject>
      <dc:subject>cash flow analysis</dc:subject>
      <dc:subject>net cash flow</dc:subject>
      <dc:subject>balance composition of cash flow</dc:subject>
      <dc:desription>In this paper deals with the cash flow analysis in agricultural companies in
        particular regions of the Slovak Republic. At present, from a company's point of view it
        is not so important to reach a profit but to have a sufficient cash to keep solvency. On
        base of column and balance cash flow composition we determine in which areas
        agricultural companies invested money and what sources they used to finance their
        activities.</dc:description>
      <dc:publisher>Faculty of Economics and Management, Czech University of Life Sciences
        Prague</dc:publisher>
      <dc:date>2009-09-30</dc:date>
      <dc:type>Paper in scientific journal</dc:type>
      <dc:identifier>http://online.agris.cz/files/2009/agris_on
        line_2009_1_gurcik_miklovicova_miklovicova.pdf</dc:identifier>
      <dc:source>Agris on-line Papers in Economics and Informatics</dc:source>
      <dc:language>eng</dc:language>
      <dc:format>application/pdf</dc:language>
      <dc:rights>Open Access</dc:language>
      <dc:coverage>World</dc:language>
     </oai_dc:dc>
    </metadata>
   </record>
  </GetRecord>
</OAI-PMH>
```

Fig. 2. Example of metadata record.

4 Conclusion

The own OAI-PMH solution for the electronic scholarly journal Agris on-line Papers in Economics and informatics was implemented in the first part of 2011 and it is available at http://oai.agris.cz. Creators and administrators of Agris on-line have prepared a general solution which can be used by other faculties' of universities' repositories. OAI-PMH application of the journal is created in PHP (PHP: Hypertext Preprocessor) with Nette Framework[1]. PHP5 is a widely-used open source general-purpose scripting language that is especially suitable for the web development (M. Achour and collective, 2011). The database for metadata library is MySQL version 5, which is the world's most popular open source database.

Metadata, which are archived in the library, are sorted to sets - for the electronic scholarly journal oai:aol (Agris on-line Papers in Economics and Informatics) is established. All the papers from the journals are available in the metadata library. Sets are optionally constructed for grouping items for the purpose of selective harvesting.

The library is managed by the Czech University of Life Sciences Prague as data provider to expose metadata to harvesters. The metadata library solution provides metadata for all sic OAI-PMH requests. All responses to the metadata library requests are well-formed XML documents. Encoding of the XML uses the UTF-8 representation of Unicode. Dates and times are uniformly encoded using ISO8601 and are expressed in UTC throughout the protocol, but the electronic scholarly journal uses date only in the YYYY-MM-DD format. Described OAI-PMH solution is prepared for content, which uses date/time too (YYYY-MM-DDTHH:MM:SSZ).

In the event of an error or an exception condition, the metadata library indicates OAI-PMH errors, distinguished from HTTP Status-Codes by including one or more error elements in the response. While one error element is sufficient to indicate the presence of an error or an exception condition, the metadata library reports all errors or exceptions that arise from the processing of the request.

Creators and administrators of Agris on-line have prepared very useful and strong application which is developed in keeping up with transnational standards OAI-PMH. The results of the OAI-PMH solution will be available for Research Program titled "Economy of the Czech Agriculture Resources and Their Efficient Use within the Framework of the Multifunctional Agri-food Systems" of the Czech Ministry of Education, Youth and Sports number VZ MSM 6046070906.

Acknowledgements. The work leading to these results has received funding from the European Commission under grant agreement nº 250525 corresponding to project VOA3R (Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment), http://voa3r.eu**.**

[1] http://nette.org/en/

References

1. Van de Sompel, H, Nelson, M. L., Lagoze, C. and Warner, S.. Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine, 10(12), 1082-9873 (2004).
2. Open Archive Initiative. The Open Archives Initiative Protocol for Metadata Harvesting. [online]. Available at http://www.openarchives.org/OAI/openarchivesprotocol.html.
3. Dublin Core Metadata Initiative. The Dublin Core Metadata Element Set. [online]. Available at http://dublincore.org/documents/dces.
4. Achour, M., Betz, F., Dovgal, A., Lopes, N., Magnusson, H., Richter, G., Seguy, D. and Vrana, J. PHP Manual. [online]. Available at http://www.php.net/manual/en/.
5. Han, M. J. Creating Metadata for Digitized Books: Implementing XML and OAI-PMH in Cataloging Workflow. Journal of Library Metadata, Volume 11, Issue 1, January 2011, Pages 19-32. ISSN 19386389.
6. Bellini, E., Deussom, M.A., Nesi, P. Assessing Open Archive OAI-PMH Implementations . DMS 2010 - Proceedings of the 16th International Conference on Distributed Multimedia Systems, 2010, Pages 153-158. ISBN 978-189170628-8.
7. Haslhofer, B., Schandl, B. Interweaving OAI-PMH Data Sources with the Linked Data Cloud. International Journal of Metadata, Semantics and Ontologies, Volume 5, Issue 1, April 2010, Pages 17-31. ISSN 17442621.
8. Van de Sompel, H. The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, February 2000, Volume 6 Number 2. ISSN 1082-9873. Available at http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html.
9. McCown, F., Liu, X., Nelson, M. L. and Zubair, M., Search Engine Coverage of the OAI-PMH Corpus, IEEE Internet Computing, vol. 10, no. 2, pp. 66-73, Mar./Apr. 2006, doi:10.1109/MIC.2006.41