Contrasting metadata quality processes and coverage in agriculture-related repositories: an experience report
George Adamides, Diane Le Henaff, Llise Ankjaer Rasmussen, Miguel - Angel Sicilia, Guy Waksman

To cite this version:
George Adamides, Diane Le Henaff, Llise Ankjaer Rasmussen, Miguel - Angel Sicilia, Guy Waksman. Contrasting metadata quality processes and coverage in agriculture-related repositories: an experience report. 8. European Federation for Information Technology in Agriculture, Food and the Environment/World Congress on Computers in Agriculture, Jul 2011, Prague, Czech Republic. hal-02750350

HAL Id: hal-02750350
https://hal.inrae.fr/hal-02750350
Submitted on 3 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Contrasting metadata quality processes and coverage in agriculture-related repositories: an experience report

P. Šimek¹, G. Adamides², D. Le Henaff³, I. A. Rasmussen⁴, M. A. Sicilia⁵ and G. Waksman⁶
¹CULS Prague, Kamýcká 129, 165 21 Prague, the Czech Republic
²ARI, P.O.Box 22016, 1516 Nicosia, Cyprus
³INRA, DV-IST, Route de St Cyr, 78026 VERSAILLES Cedex, France
⁴ICROFS, Blichers Allé 20, Post-box 50 DK-8830 Tjele Denmark
⁵UAH, Ctra. de Barcelona, Km. 33,600, 28871 Alcalá de Henares, Spain
⁶ACTA & ACTA Informatique,149 rue de Bercy - FR - 75595 PARIS Cedex 12, France
simek@pef.czu.cz

Abstract

Scholarly content is nowadays exposed through a diversity of systems implementing different workflows and metadata schemes. The integration of such systems thus requires a previous phase of analysis that uncovers the differences as a first step towards their integration. This paper describes an experience report on doing such an analysis for five different digital collections of open access scholarly content related to agriculture. Taking as a common schema the Dublin Core metadata element set, both the use of different metadata elements and also the wider context in which metadata is created and quality controlled are examined.

Keywords: Metadata, OAI-PMH, repository, open access

Introduction

Open access to scientific/scholarly content removes the access barriers related to copyright retention to the outcomes of research. In COM(2007)56 (Commission of the European Communities, 2007), the Council of the EU highlighted that access, dissemination and preservation of scientific (including both publications and research data) are crucial elements of the development of the European Research Area, in order to accelerate EU innovation and competition. Open access publishing has grown steadily in the past recent years. This can be partly attributed to the fact that traditional journal subscriptions are getting more and more expensive. The increasing cost of academic journals makes it difficult for scholars from less wealthy regions or institutions to get the information they need to conduct their own research. Open access is an important publishing alternative that was developed to help to sort out this problem (Stranack, K, 2008).

Institutional repositories and open journal systems are four examples of repositories of scholarly content that use metadata for the description of their digital collections. These different systems can be integrated into centralized repositories by a process of mapping the metadata schemas used and harvesting them into a single infrastructure. Examples of this kind of integrated systems are Europeana⁴ and LRE4Schools. In the domain of agriculture and aquaculture, the VOA3R (Virtual Open Access Agriculture & Aquaculture Repository) platform aims at providing that kind of integrated service for the retrieval of relevant open

⁴http://www.europeana.eu/portal/
content and data. From a technical perspective, these integrated services have the form of a federated repository, having a metadata repository as the central point around which search, navigation and services are provided. Protocols as the OAI-PMH are mainstream means for achieving the technical integration.

However, the integration of digital collection following the abovementioned approach requires considering the heterogeneity of the local repositories beforehand. Concretely, not all the existing systems are using metadata schemas consistently, and the processes for the creation and quality control of the metadata they have in place are also widely heterogeneous. This paper reports the experience in contrasting five different digital collections in the domain of agriculture as an illustration of the difficulties that arise when approaching the integration of disparate systems.

**Data and methods**

This section provides the description and context of the five digital collections compared.

**Organic Eprints**
Research results regarding organic food and farming mainly from Europe but also from the rest of the world have become easily accessible. The Open Access archive Organic Eprints (orgprints.org) has developed since the start in 2002 so that it now includes more than 10,000 items, has 15,000 registered users and 175,000 visits per month. The full text documents are pdf's, word-files, power-point-presentation and many other types of files. The archive is open for all to use and registered users can deposit their research publications from refereed journals as well as non-refereed sources. Organisations, research facilities, research programmes and projects are also presented in the archive. Organic Eprints is running on Linux and other open source software. It is based on the EPrints 3 archive-creating software, which is available for free from eprints.org. The archive is based on GNU Eprints archive-creating software. GNU Eprints generates archives that are compliant with the Open Archives Protocol for Metadata Harvesting.

**The ARI Repository**
The Agricultural Research Institute (ARI) Repository is hosted at http://library.ari.gov.cy/agris, and it is built upon WebAGRIS version 2.0. WebAGRIS is realized by AGRIS/CARIS and Documentation Group, GILW, FAO. The interface is based on HTML forms, and has been implemented as a CGI program. The program is invoked by the XITAMI Web server process. The ARI repository provides bibliographic references and full document access to publications of the Agricultural Research Institute (http://www.ari.gov.cy). These records are sent every year to AGRIS (International System for Agricultural Science and Technology) – using the AGRIS AP for data exchange. The Global Database is maintained by FAO, while the local database is maintained by the Agricultural Research Institute which is the national centre for AGRIS in Cyprus. There are more than 650 records in the ARI Repository and all records are in Agris AP format.

**The Agris on-line Papers in Economics and Informatics**
The international journal AGRIS on-line Papers in Economics and Informatics is a scholarly open access, blind peer-reviewed, interdisciplinary, and fully refereed scientific journal. The
journal is published quarterly by the Faculty of Economics and Management, Czech University of Life Sciences Prague. 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. There are more than 60 scholarly papers in pdf format at the journal repository and papers are inserted quarterly. The journal is available at http://online.agris.cz.

The EFITA repository

The efita.net web site has been created by ACTA Informatique (Paris, France) in 2003 and it presents the archives of the week efita newsletter as well as these of the efita congresses that were held every two years since 1999 (See: www.efita.net).

The efita.net web site is managed thanks to an “in-house” solution based on the use of a relational data base management system named Firebird. Number of items (records) in EFITA db is around 2,000. ACTA Informatique is working on the migration of the efita.net web site under the (Agri) Drupal CMS that enables to easily produce RDF format files (Repository data file) that should expose on the web the metadata extracted for the efita.net records.

The Drupal CMS solution, being promoted by FAO ACTA Informatique hopes to benefit from experiences of other organisations in the management of metadata exposed on the web for further harvesting (See: aims.fao.org/website/AgriDrupal/sub2).

ProdINRA, the INRA institutional open archive

The French National Institute for Agricultural Research (INRA) is the world’s second-largest research institute in the field of agriculture, food and nutrition and environment. Its international visibility and recognition are a result of excellence and the widespread dissemination of its research work. In 2005, INRA demonstrated its support for open access by signing the Berlin declaration and launching an institutional open archive called ProdInra. Since that year, INRA has been overhauling its information system in order to set up indicators for steering agricultural research, but also to provide multiple services to INRA personnel as well as the agricultural scientific community at large. In its four-year strategic contract, INRA restated its desire to develop its international activities. Promoting INRA research and highlighting its scientific and technical production internationally serve to generate collaborations and contribute to INRA’s image as a major player in international agricultural research.

The ProdInra staff has therefore rolled out collaboration strategies in order to promote INRA’s scientific production by having it referenced in numerous major document databases. INRA has then become a VOA3R partner. In order to make the partnership the most efficient, ProdInra is exposing the 6,137 records through an OAI repository5. All these records have been validated by an information specialist during the review process.

5 URL of the OAI server: http://www.prodinra.inra.fr/prodinra/sdx/oai/pinra/notices/oairepo-fulltext
Requirements of the integration

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 et al., 2004). There are two classes of participants in the OAI-PMH framework:

- Data Providers administer systems that support the OAI-PMH as a means of exposing metadata; and

- Service Providers use metadata harvested via the OAI-PMH as a basis for building value-added services (Open Archive Initiative, 2008).

![OAI-PMH functionality.](image)

The Data Providers administer their local repositories (network accessible server) with local services for data access. They must provide software application, which can process the OAI-PMH requests. The Service Provider uses metadata harvester. The metadata harvester is a client application that issues OAI-PMH requests. Harvested metadata are saved to the central repository. The Service Provider provides data access via central services. The XML records harvested need to have a consistent format for their consumption into the central system. Also, provenance information is required in the metadata, as there may be differences on quality and breadth of coverage depending on the metadata workflows in place at each system. These are described for the systems analysed in the following section.

Results and discussion

The integrated repositories include very different contents, but the metadata of the contents are similar. A few repositories used basic Dublin Core standard, which includes 15 recommended elements (Dublin Core Metadata Initiative, 2010). However the repositories do not use all 15 elements.
Figure 2. Metadata in the EFITA, Agris on-line Papers in Economics and Informatics, Organic Eprints and ProdInra.

The ARI repository includes following information, which is used to create an AGRIS record: ARN (Agris Resource Number), Data Entry Status (temporary or completed), Type of Publication, Bibliographic Level, Literary Indicator, Medium (form), Agris record, Author’s name, Affiliation, E-mail, Role, Corporate author, English title, French title, Spanish title, Other language title, Conference name, Conference number, Conference place, Conference date, Report/patent number, Collation, Language of text, Summary statement, Notes, URL, Internal access only, File extension, Language, Serial Main
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 from all OAI-PMH requests, so:

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

**Conclusion**

Some repositories have used open source software like DSpace, Drupal (AgriDrupal) or Eprints while some others use their own special solution. In any case, different systems are making use of schemas and encoding patterns differently and they are managing metadata using different workflows.

All local repositories provide or will provide an application for metadata harvesting, which can process the OAI-PMH requests and support selective harvesting with the from and until arguments expressed at day granularity. Metadata harvesters may use date stamps to harvest only those records that were created, deleted, or modified within a specified date range too. The value of date stamps is in UTC date time format.

Federated repositories rely in a distributed approach to content management, with each source system being responsible of storing the actual content, and using metadata harvesting to the central metadata repository. This requires a previous step of analysis to identify the differences and understand their implications for the services provided for the integrated collection.

**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.
References


DUBLIN CORE METADATA INITIATIVE. The Dublin Core Metadata Element Set. [online]. Available at http://dublincore.org/documents/dces/.