Reorienting open repositories to the challenges of the Semantic Web: Experiences from FAO's contribution to the resource processing and discovery cycle in repositories in the agricultural domain

Imma Subirats*, Thembani Malapela*, Sarah Dister*, Marcia Zeng**, Marc Gooaverts***, Valeria Pesce****, Yves Jaques*, Stefano Anibaldi*, Johannes Keizer*

* FAO of the UN (Italy), ** Kent State University (USA), *** Hasselt University Library (Belgium), *** Global Forum on Agricultural Research (Italy

Abstract. The use of widely-used metadata standards is essential to guarantee the visibility and retrieval of documents stored in open repositories. Attention should be paid to the creation and exchange of meaningful metadata to enhance interoperability amongst repositories and provide value added services. Since 2005 the Food and Agriculture Organization of the United Nations (FAO) provides the agricultural information management community with standards, services and tools to assist open repositories in benefiting from the advantages offered by Semantic Web publishing. This paper presents the work that FAO carries out in recommending standards for the encoding and exchange of metadata while also reviewing techniques to help navigate within open repositories and services. It talks about how to improve the visibility of repository content and explains the benefits of integrating subject vocabulary tools expressed in SKOS. It concludes with a presentation of use cases integrating these recommendations into DSpace and Drupal customizations.

Keywords: Open Access ;Open Repositories; Metadata; Repository Interoperability, AGROVOC; AgriOcean DSpace; AgriDrupal; WebAGRIS; LODE-BD; Linked Data; AIMS; Semantic Web; AgMES; AGRIS AP

1 Introduction.

The Open Access movement satisfies two broad intertwining goals: firstly, facilitating the online archiving of digital documents (in most cases peer reviewed post prints) and making them freely accessible through an OAI compliant repository (Green route); secondly, sustaining open access journals by depositing articles online upon publication (Gold route). The acceptance and growth of this model and its hybrids in the scholarly communication process has seen an increase in the number of open repositories available online; for instance, OpenDoar [1] reported 2,211 repositories registered by September, 2012.

Current technological changes especially in the Semantic Web dictates that open repositories should *not only publish local content globally,* but also offer additional *values to researchers* by harnessing participation from a broad community of data providers (*interoperability*). In this way, open repositories are poised to increase the role they play within the scholarly communication process. However, certain fundamentals have to be met if open repositories are to remain visible.

The Semantic Web has further facilitated value addition to research outputs through automatic discovery, linking and analysis. Linked Data is the set of best practices for publishing and connecting structured data on the web. Its main objective is to liberate data from silos that are framed by proprietary database schemas by following the four principles, as defined by Tim Berners-Lee [2] in 2006.

In the agricultural domain, FAO has been providing support to agricultural information communities to build and maintain open repositories that conform to recommended metadata standards. In this vein, this paper presents the role of the Agriculture Information Management Standards (AIMS) team in *re-orienting* repositories to the current demands of the Semantic Web, through (a) AIMS set of recommendations to open repositories; and (b) providing FAO's experiences and use cases in implementing these recommendations.

2 Literature Review.

The major goal of digital repositories is to facilitate access to their contents. Swan and Carr aptly re-state that,

> "Repositories should be one of the institution's web based tools that take research into places that have not been reached before. One important issue … is that the primary reason for establishing a digital repository is to increase the visibility of the institution's research output by making it available on Open Access." [3]

Visibility has been defined in the context of repositories to mean the number of external links received by a repository from external sites [4], [5]. The total visits made to a repository contents by links from search engines and other databases is used to measure visibility. *The Ranking Web of World Repositories* was started with the aim to improve visibility of open repositories and to promote good practices in their publication [6]. The methodology employed by the *Ranking of the Web of World Repositories* includes the following parameters, Size; *Visibility*; Rich files and Scholar (The total number of papers in Google Scholar for a 5-year period 2007-2011)

Most repositories strive for global visibility and to fully expose their contents. [7],[8],[9]. Yet a recent study by Artlitsch and O'Brien [10] established that most repositories are invisible, for example Google Scholar had difficulty in indexing the contents of institutional repositories, and Artlitsch and O'Brien hypothesized that most repositories use Dublin core, which cannot express bibliographic citation data adequately for academic papers. During this study, experimental metadata transformation projects were implemented at Utah and were successful in achieving a greater than 90% indexing ratio. It is clear that the quality of metadata records stored in repositories assures greater visibility.

Still, when different metadata standards and schemas are used across repositories this creates challenges in achieving interoperability [11] and Haslhofer and Klas [12] proposed metadata integration to solve this. However, Park and Lu [13] discovered that even in the use of a common metadata standard there was a divergence in what local metadata guidelines contained and what they represented. This was found to be a potential hindrance to sharable metadata across repositories. Therefore, attention to the standardization of metadata at individual field level within a resource is important if the efficient retrieval of documents stored in open repository is to be achieved.

The use of vocabulary control has also been proven to be effective in retrieval of information in electronic environments [14]. In the context of the Semantic Web it has been noted [15] that the use of controlled vocabularies is useful in the retrieval and discovery of resources tagged with repository concepts. Gray *et al* [16] phrased it this way;

> "Using SKOS as a representation for a vocabulary provides a unique identifier to tag resources with, and enables vocabulary aware applications to enhance…the exploitation of relationships between concepts in the vocabulary…..vocabulary aware applications can

> benefit from improvements in both precision and recall, for example when searching for bibliographic or science data."

When repositories use controlled vocabularies in indexing their content great success in resource discovery improves and also facilitates easier resource sharing amongst repositories.

3 Recommendations to Open Repositories.

If repositories are to remain open and accessible in the Web of data, they must ensure that:

i. their content is stable (browsable, searchable, discoverable, and readable by both machines and humans);
ii. they use appropriate metadata standards to improve exchange across data silos;
iii. they use controlled vocabularies and ensure that these are integrated within document repository management systems (essential if these vocabularies are in themselves Linked Open Data!).

Therefore, with regards to item ii. and iii. stated above, AIMS recommends that repository managers should use *Linked Open Data Enabled Bibliographic Metadata (LODE-BD) recommendations*[17] in deciding which metadata properties to use. Whereas with regards to the use of controlled vocabularies, agricultural repositories are encouraged to use the AGROVOC to describe the contents of their repositories. With the launch of the AGROVOC linked open data, repositories can simply link their resources to AGROVOC and this model has been successfully applied elsewhere [16] and within the agricultural domain [18]. The following subsections will provide an elaborate description of this model.

3.1 The key step towards semantic interoperability: assuring quality in metadata creation.

Metadata in repositories serve both an administrative role during the submission process and a technical role of resource description for resource discovery by a broad audience. If repositories are to operate across administrative and disciplinary boundaries, and are to be relevant in the Semantic Web, they should guarantee resource-level accessibility. Content description and indexing through standardized metadata, when applied to both syntax and semantics, becomes the basis of efficient visible repository to which value-added services can also be harnessed.

The AGRIS (International System for Agricultural Science and Technology) Network[19], is an international information system for sharing access to agricultural science and technology information created in 1974. It is a collaborative system which includes more than 100 national, international and intergovernmental centres with a goal to facilitate information exchange of literature dealing with all aspects of agriculture. As a result, the AGRIS Network contributes to the AGRIS Database, a content aggregator with 2.9 million bibliographical records on agricultural science and technology, maintained by FAO.

Since 2005 the AGRIS Application Profile (AP) [20] has been used as a metadata schema for the submission of agricultural information metadata to AGRIS, superseding the earlier version, AGRIN [21]. The AGRIS AP uses metadata elements from Dublin Core (DC), Australian Government Locator Service Metadata (AGLS) and Agricultural Metadata Element Set (AgMES) , developed by FAO in 2003. The AGRIS AP enforces a minimum level of quality and the use of controlled vocabularies by mandating four required elements and promoting the use of agriculture-specific thesauri such as AGROVOC [22]. The new demands of the Semantic Web and its open-world assumptions have revealed the limitations of the AGRIS AP. It seemed to be too rigid in its encoding requirements while at the same time promoting a number of properties that are too obscure for an open-world approach. It is thus not been able to guarantee interoperability among data providers and services, particularly beyond the agricultural information management community.

In 2011, FAO re-oriented its approach by providing a set of recommendations with a full range of options for metadata encoding from which bibliographic content providers could choose according to their development stages, internal data structures, and the reality of their current practices. The recommendations allow any content provider to encode bibliographic data[1] using properties from standardized namespaces, to use well-established authority data and controlled vocabularies available as linked data in agriculture and to publish data in RDF. The recommendations encourage data providers to adopt good encoding strategies to facilitate the exchange of bibliographic metadata. These recommendations are referred to as *Linked Open Data Enabled Bibliographic Metadata* [17] (LOBE-BD) version 2.0. LOBE-BD assists repository managers in four key

[1] An instance of bibliographic resource includes articles, monographs, theses, paper, material presentation, research report, learning object, etc. - printed or electronic format

questions: (a)What **kinds of entities and relationships** are involved in bibliographic resource descriptions? (b) What **properties** should be considered for publishing meaningful/useful Linked Open Data-ready bibliographic data? (c) What **metadata standards** should be used for preparing Linked Open Data-ready bibliographic data? (d) What **metadata terms** are appropriate in any given property for producing Linked Open Data-ready bibliographic data from a local database?

Although LODE-BD focuses on the exchange of data, it also contains recommendations about the minimal set of metadata properties, and syntax encoding rules, controlled vocabularies and authority data, necessary to produce, manage and exchange meaningful bibliographic metadata. LOBE-BD recommendations provides practical decision trees in selecting properties in its nine groups. The decision trees are arranged in a flow chart which highlights decision points and gives a step-by-step solution to a given metadata encoding. The Figure 1 below shows the example for the decision tree for Title information.

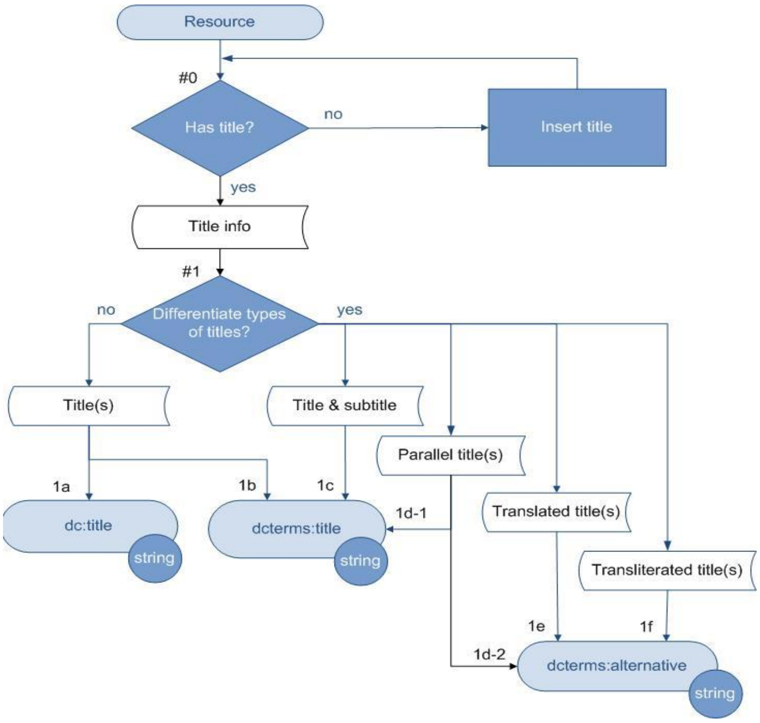

Fig. 1. LOBE-BD Decision tree – Title Information

3.2 Aids to navigation and visibility of repository contents

Most repositories have adopted the use of URLs in identifying their resources as a first step towards creating visibility of their holdings. However, differences that arise due to geographic, cultural, domain specific environments even amongst repositories with the same or similar collection scopes, still inhibit individual resource visibility.

Subject vocabularies (words or phrases taken from standardized, organised knowledge structures) should be employed to resolve indexing problems such as plurals, spelling variants, synonyms and homonyms (same spelling representing two different concepts, e.g. *blood* vessel / *fishing* vessel). In the context of Semantic Web such subject vocabularies are expressed as a concept scheme using SKOS (Simple Knowledge Organisation System) and integrated within document management or content management systems. The use of subject vocabularies guarantees meaningful metadata while also enhancing the quality of the interoperability and effectiveness of information exchange among data providers, thus facilitating the re-usage of data by other repositories/services and in the process adding value to the local researcher.

The AGROVOC [23] thesaurus contains more than 40,000 concepts in up to 22 languages covering topics related to food, nutrition, agriculture, fisheries, forestry, environment and other related domains. AGROVOC is a thesaurus expressed as a concept scheme using SKOS and this conversion from a relational database has provided added semantics value to term relationships. Therefore, current structure of AGROVOC concept scheme provides three levels of presentation. These three levels are: A) **CONCEPTS** – refers to the abstract meaning and often identified using URIs, for example maize in the sense of a cereal identified by Concept12332;B) **TERMS** - are language-specific lexical forms attached to concepts, for example maize, maïs, 玉米, ข้าวโพด, or corn, C) **TERM VARIANTS** - are the range of forms that can occur for each term such as spelling variants, singular or plural, for example organization or organisation, cow or cows.

In partnership with MIMOS Behard [24], the AGROVOC thesaurus is published as a Linked Data aligned more than ten other knowledge organization systems. The additional value that linking AGROVOC to other vocabularies provides is that ***data repositories attached to those vocabularies become discoverable***. This is a very simple classic case of exposing repository contents automatically across datasets through AGROVOC indexing.

4 Use cases in integrating information management standards in selected Information Management (IM) tools.

Three open source management tools have been customized to facilitate the use of standards for the creation, management and exchange of metadata.

4.1 WebAGRIS

WebAGRIS [21] is an information management system for the creation and dissemination of AGRIS AP metadata based on WWW-ISIS software[25] and customized by the Institute for Computer and Information Engineering in Poland with the support of FAO [26]. Despite the obsolescence of the technology used by WebAGRIS, during the last 10 years it has been the most widely-used information management tool within the AGRIS Network. This is due to the fact that WebAGRIS does not require a complex technical infrastructure for its maintenance, a key selection point for many developing countries. WebAGRIS provides functionalities like protected access for creation and update of metadata and export of AGRIS AP records, authority data creation and maintenance (e.g. AGROVOC Thesaurus built-in), user friendly retrieval and AGROVOC Thesaurus based search. WebAGRIS can be used in a LAN or WAN, so multiple nodes may contribute to a centralized instance of WebAGRIS, simply via an IP. In 2012 the FAO AIMS team has stopped supporting new developments on WebAGRIS, and discourage new users to install it . However support to existing users will continue.

4.2 AgriOcean DSpace

DSpace is an open source and freely available software conceived for the setting up and management of open repositories. DSpace focuses on managing and preserving digital content. It is based on a solid community of DSpace users and developers. It is possible to customize it and extend it. In 2009 the FAO AIMS team, in collaboration with Hasselt University, the Institute of Biology of the Southern Seas (IBSS) and UNESCO-IOC/IODE, proceeded with a customization of DSpace, AgriOcean Dspace

(AOD), based on specific information management standards widely used in the agricultural, aquatic and marine sciences.

AOD supports the use of rich metadata element set and subject vocabularies/authority control for the description of any type of information, like journal contributions, books, conference contribution, research report, working papers, theses or other like preprints. The main features introduced by AOD are the following: i)exposure of records through the OAI-PMH protocol supporting metadata formats like AGRIS AP and MODS [27] ;ii) indexing with ASFA and AGROVOC terms; iii)authority control features for journal title; iv) submission base on type of document; v)easy to install version for Windows; vi) up-to-date lay-out: personalizable standard vii)batch import for AGRIS AP, MODS and EndNote.

AOD is based on the out-of-the-box DSpace, which its main features and functionalities are: self-archiving and submission process, different submission workflows, management of digital objects, variety of digital format and content types are supported, two levels of search, persistent identifiers (handle),long-term physical storage and OAI compliancy and RSS exposure. AOD is available in source code or with a Windows installer designed specifically to make it easy to install for organizations with limited IT support. AOD is currently used by Oceandocs [28], the Institute of Biology of the Southern Seas, Ukraine (IBSS) [29], Central and Eastern European Marine Repository (CEEMaR) [30] and the Ministry of Agriculture (Peru)[31], and is under testing by other 13 institutions.

4.3 AgriDrupal

In setting up repositories, agricultural institutions have often faced the following demands in the selection of appropriate software tools: the need to integrate a repository search and browse interface within their website, the need to implement custom content models, or custom metadata models, and,the need to be able to exchange information with other systems and participate in other networks[32]

In 2009 the FAO AIMS team initiated the project AgriDrupal [33] as a *suite of solutions* for agricultural information management and dissemination, built on the Drupal [34] platform, with special functionalities for repository management.In 2010, FAO piloted an AgriDrupal installation at the National Food Policy Capacity Strengthening Programme (NFPCSP) [35] in collaboration with the Ministry of Food and Disaster Management

(MoFDM) in Bangladesh; with financial support from the European Union and the United States Agency for International Development.

The pilot made it apparent that the AgriDrupal tool was quite appropriate for managing both the electronic documentation centre and a website adopting standards that FAO had also supported [36]. AgriDupal has since been offered to agricultural information managers as an integrated solution to manage different types of information such as organizations, expert profiles, news, jobs, events, feeds, web pages, blog entries or forum topics. It has advanced features for managing Open Access document repositories in compliance with widely adopted library standards. Each AgriDrupal installation now comes with the following added-value features: i) import and export functionalities using the AGRIS-AP XML format for bibliographic records and extended RSS for other types of records; ii)ability to index any content with AGROVOC terms; iii) exposure of bibliographic records through the OAI-PMH protocol supporting two metadata formats (Dublin Core and AGRIS AP); iv) support for implementing additional metadata standards; v) all the core Drupal Content Management features for advanced management of any contents and customization of the look and feel. The AgriDrupal installation has been used also by the Ghana Agricultural Information Network System (GAINS) portal and recently by the ZAR4DIN [37] national portal in Zambia in managing their website as well as their document repositories via a single interface.

5 Conclusions.

In this paper we have advocated that repositories need to strive for continuous visibility and guarantee interoperability. It has been established that most repository are invisible when searched by search engines and the semantic web threatens to render such resources further invisible in the future if they remain in their present form. In order to reorient open repositories to the demands of the semantic web, we proposed two basic interventions, the first is that repositories should adopt widely-used metadata standards for the description of information objects. Secondly, repositories should use controlled subject vocabularies which are expressed as a concept scheme and are in Simple Knowledge Organisation System (SKOS) in indexing their contents. The FAO AIMS team, therefore, recommends that AGROVOC Thesaurus as linked data is a good subject vocabulary for indexing contents for repositories in the agricultural domain. Practical examples were offered in the agricultural information management domain highlighting how the AgriOcean DSpace and AgriDrupal software(s) have integrated these recommendations ; these were also presented as open repositories use cases. Despite this model, there still remain an opportunity for further research into how open repositories can be migrated into the semantic web by having them published as Linked Open Data

6 References.

1. OpenDoar http://www.opendoar.org/index.html Last accessed: July 2012.
2. Tim Berners-Lee (2009) Linked Data. http://www.w3.org/DesignIssues/LinkedData. Last accessed: July 2012
3. Swan, A., and Carr,L (2008) Institutions, their repositories and the web. Serials Review. Vol. 34 (1), 31-35.
4. Drewry, J (2007). Google Scholar, windows live academic search, and beyond : a study of new tools and changing habits in ARL libraries. A Master's Thesis. http://ils.unc.edu/MSpapers/3310.pdf Last accessed September 2012.
5. Aguillo, I., Ortega,J., Ferriandez, M., and Utrilla, A (2010) Indicators for a webometric ranking of open access repositories. Scientometrics. Vol.82 (3), 447-486.
6. Ranking Web of Repositories website. http://repositories.webometrics.info/en. Last accessed: September 2012
7. Abrizah, A., Noorhidawati, A and Kiran, K (2010) Global visibility of Asian Universitie's Open Access institutional repositories. Malaysian Journal of Library and Information Science. Vol.15 (3) 53-73.
8. Banier,J and Perciali, I (2008) The institutional repository rediscovered; what can a university do for Open Access publishing? Serials Review. 34 (1), 21-26.
9. Mercer, H., Koeing,J., McGeachin,R., and Tucker,S. (2011) Structure, features, and faculty content in ARL member repositories. The Journal of Academic Librarianship. Vol. 37 (40), 333-342.
10. Arlitsch, K., and O'Brien, P. S. (2012). Invisible institutional repositories: addressing the low indexing ratios of IRs in Google Scholar. Library Hi Tech, 30(1), 60–81.
11. Ochoa, X., and Duval, E (2009) Automatic evaluation of metadata quality in digital repositories. International Journal on Digital Libraries. 10(2-3), 67-91.
12. Haslhofer, B., and Klas, W. (2010). A survey of techniques for achieving metadata interoperability. ACM Computing Surveys, 42(2), 1–37.
13. Park, J., and Lu, C (2008). An analysis of seven metadata creation guidelines; issues and implications. A paper presented at 2008 Annual Electronic Resources and Libraries Conference, Atlanta, Georgia. March 18-21, 2008.
14. Michael, G (2004) Authority control in the context of bibliographic control in the electronic environment. Cataloguing and Classification Quartely.38 (3-40), 11-22.
15. Weller, K (2010) Knowledge representation in the social semantic web. New York: Walter de Gruyer.
16. Gray,A., Gray,N., Hall,C and Ounis, A (2010) Finding the right term: retrieving and exploring semantic concepts in astronomical vocabularies. Information processing and Management. Vol.46 (4) 470-478.

17. Subirats,I and Zeng,M (2012)Linked Open Data Enabled Bibliographic Metadata (LODE-BD) version 2.0 http://aims.fao.org/standards/lode-bd Last accessed: September 2012.
18. Lukose,D (2012) World-wide semantic web of agriculture knowledge. Journal of Integrative Agriculture. Vol. 11(5) 769-774.
19. Knowledge and information sharing through the AGRIS Network http://agris.fao.org/knowledge-and-information-sharing-through-agris-network Last accessed: July 2012
20. FAO(2005)The AGRIS Application Profile for the international information system on agricultural sciences and technology guidelines on best practices for Information Object Description http://www.fao.org/docrep/008/ae909e/ae909e00.htm Last accessed: July 2012
21. Onyancha, I ., Weinheimer, J ., Salokhe, G., Katz, S., and Keizer, J. (2004). Metadata exchange without pain: the AGRIS-AP to harvest and exchange quality metadata http://dcpapers.dublincore.org/index.php/pubs/article/download/774/770 Last accessed: July 2012
22. Subirats, l,. Onyancha, I., Salokhe, G., Keizer, J.(2008) Towards an architecture for open archive networks in Agricultural Sciences and Technology ftp://ftp.fao.org/docrep/fao/009/ah766e/ah766e00.pdf Last accessed: July 2012.
23. AGROVOC Homepage http://aims.fao.org/standards/agrovoc/about Last accessed: July 2012
24. MIMOS Behard Homepage http://www.mimos.my/ Last accessed: July 2012
25. UNESCO. WWW/ISIS – Technical Reference Manual, v. 5.1.1 / 05-05-02, Warsaw/Rome, May 2005.
26. Rybinski,H , Kaloyanova, Sand Katz,S. (2006) WWW-ISIS: a result of a close cooperation between FAO-GIL and ICIE ftp://ftp.fao.org/docrep/fao/010/ai162e/ai162e00.pdf Last accessed: July 2012
27. Metadata Object Description Schema (MODS) http://www.loc.gov/standards/mods/ Last accessed: July 2012
28. OceanDocs http://www.oceandocs.org/ Last accessed July 2012
29. IBSS Institutional Repository http://repository.ibss.org.ua/dspace/ Last accessed July 2012
30. Central and Eastern European Marine Repository http://repository.ibss.org.ua/dspace/ Last accessed July 2012
31. Ministry of Agriculture (Peru) http://www.minag.gob.pe/portal/ Last accessed February 2012
32. Pesce, V., Subirats, I., Picarella, A., Keizer, K.(2011) AgriDrupal : repository management integrated into a content management system,. A paper presented at Open Repositories Conference 2011,Austin (US),8-10 June 2011.
33. AgriDrupal Homepage http://aims.fao.org/tools/agridrupal Last accessed: July 2012
34. Drupal Open Source Content Management System http://drupal.org/ Last accessed: July 2012
35. NFPCSP Homepage http://www.nfpcsp.org/agridrupal/ Last accessed: July 2012
36. AgriDrupal at NFPCSP http://aims.fao.org/advice-and-capacity-development/open-access/fpmu Last accessed: July 2012.
37. ZAR4DIN Homepage http://zar4din.org/ Last accessed: July 2012