Metadata Approaches for Shareable and LOD-enabled Bibliographic Data from Open Repositories

Imma Subirats
FAO of the UN, Italy
Imma.Subirats@fao.org

Marcia Lei Zeng
Kent State University, USA
mzeng@kent.edu

Johannes Keizer
FAO of the UN, Italy
Johannes.Keizer@fao.org

Keywords: interoperable metadata; LOD-enabled metadata; open bibliographic data; metadata mapping approaches; methodology

This poster presents the processes and paths of the authors who have recently prepared a report on descriptive metadata encoding recommendations for an European project, VOA3R (Virtual Open Access in Agriculture and Aquaculture Repository), which aims to deploy a virtual entry-point for exchanging and augmenting open bibliographic data, and thus improve the dissemination of research results in agriculture and aquaculture via open access. Specifically, our task was to prepare a report with a suggested title of "Recommendations for the Content Population of the VOA3R Service Provider". Since the VOA3R Federation consists of 17 institutions from 13 countries contributing proximately two million bibliographic records to eight open repositories, the immediate need was for the analysis of the number and characteristics of the open access documents that will be accessible from VOA3R. Following this task, the next step was to propose encoding recommendations for the exchange of metadata between data providers and the VOA3R platform. Along with the wave of the Linked Open Data (LOD) movement, the VOA3R project required that the recommendations should also be suitable for encoding with consideration to Linked Open Data.

Experiences and lessons learned from each of the steps will be shared through the poster, including:

1. Collecting questionnaires and data dictionaries

At the first phase, data was collected from the eight data providers. This data consists of three major parts: i) answers to a questionnaire that would provide general description of the repositories and characteristics of internal data structures; ii) data dictionaries, and iii) sample metadata records.

2. Clustering the elements used

The focus of the data analysis was on the internal structures of the participating repositories. The metadata elements used by the eight providers are very different not only in terms of the elements, labels and scope, but also in the overall data structures. For example, some providers employ a flat structure and some others use a hierarchical structure. The implementations of authority control for responsible bodies and for subject indexing are inconsistent. Description levels are also diverse, for example, on the analytical level or aggregate level for proceedings and collective volumes. Using the terminology of FRBR (Functional Requirements for Bibliographic Records), in these data repositories, the treatment of *work*, *expression*, and *manifestation* are not consistent in either type of entities or relationships between entities. Metadata elements that share the same or similar labels may correspond to different entities. Administrative metadata is included in some repository's data dictionaries. As a result, we used a spreadsheet to align all of them on over 120 rows in a table and made loose clusters.

3. Identifying chunks and properties

In identifying the properties to be included in the recommendations, both top-down and bottom-up approaches were used. The top-down approach was to utilize a conceptual model for sharing the common understanding of the important entities and relationships for bibliographic data. The conceptual model was built on a FRBR-based model, previously developed by the FAO AIMS (Agricultural Information Management Standards) Group, with significant extension and reconsideration for this project. A general model was created at a high level of abstraction and an implementation-level model gave details of possible relationship types between the entities

Resource, *Agent*, and *Thema*. Major relation types were identified between a resource instance and the agent(s) that are responsible for the creation of the content and the dissemination of the resource, as well as between a resource instance and the thema(s) (subjects or topics) that the resource's content is about.[1]

The bottom-up approach was to analyse the data dictionaries and sample records provided by the data providers participating in the VOA3R project, from which we identified two dozen common properties for describing bibliographic resources and grounded them into nine groups. These include title information, responsible body, physical characteristics, location, description of content, subject, intellectual property, and usage, as well as an additional group for properties describing relations between bibliographic resources or between agents.[2]

4. Deciding the basic approach

Making a decision of the basic approach was a milestone in the process of preparing the final recommendation report. The original plan was to develop two application profiles. The first one would be a Dublin Core (DC)-based application profile, which would not only allow general harvesting but also enable connecting to the LOD-enabled bibliographic data. The second one would be an application profile based on MODS (Metadata Object Description Schema), which would be used by those repositories that have rich bibliographic data. The MODS-based data would need to be converted to the DC-based through a crosswalk if moving into the LOD cloud. The benefit of such an approach would be to reuse or follow some existing application profiles and crosswalks developed by various metadata communities while also ensuring that some repositories will be able to preserve more detailed data structures. In this case, the eight data providers would use the proposed VOA3R application profile(s). The disadvantage of this approach would be the additional mapping/converting processes from the original local data to the VOA3R data silos before moving into the LOD cloud.

FAO's experience with the AGRIS Application Profile (AP) was also considered. Whereas the AGRIS AP had been a useful exchange format within the AGRIS network in the last five years, the limitation of accepting data that are semantically richer or less rich is a constraint. Further the revision of the DCMI Metadata Terms (dcterms) from October 2010 was studied and considered in this project.

Out of all these considerations, a new idea in the VOA3R data harvesting approach emerged: In addition to the original plan, a different strategic approach would be to create a set of recommendations with a full range of options for metadata encoding that data providers can choose from to meet the needs at their development stages, internal data structures, and the reality of practices. The recommendations would allow any data provider to encode bibliographic data using properties from widely-used standard namespaces, to use well-established authority data and controlled vocabularies that are available as linked data in agriculture and aquaculture, to publish data in RDF triples and to submit the dataset to VOA3R. In doing so, VOA3R would act both as a service provider that enhances the dissemination channel and accessibility of open access documents and as a service that promotes the exchange and publication of bibliographic data in RDF, so as to facilitate the use of Linked Data in agriculture and aquaculture.

5. Analyzing scenarios and developing decision trees

By the time the report was prepared at the end of 2010, certain authority files and controlled vocabularies had been published or in planning to be published as Linked Data. FAO's AGROVOC which has been a major thesaurus used by the majority of VOA3R participants was also ready to be released as Linked Data. (It was released as Linked Data in April 2011.) In addition, FAO AIMS authority data – the Journal Authority Data (JAD) Collection -- was already modelled as concepts and labels. On the other hand, some repositories participating in the VOA3R project had not or only partially implemented authority control. Considering the wide range of needs, we created a set of recommended decision-making trees for common properties used in describing a bibliographic resource. Different scenarios reflecting the situations and stages of a data provider were built into the decision trees represented by flowcharts and tables.[3] Eighteen flowcharts were drawn with standardized symbols to portray steps and processes involved in decision making. These flowcharts were designed to facilitate the selection of the appropriate strategies

[1] See detail at: http://aims.fao.org/lode/bd/core-entities
[2] See detail at: http://aims.fao.org/lode/bd/properties
[3] See detail starting at http://aims.fao.org/lode/bd/decision-trees

adjustable to data providers according to their situations, while all moving towards the goal of data exchange and reuse. Starting from the property that describes a resource instance, a flowchart presents decision points and gives a step-by-step solution to a given problem of metadata encoding. At the end of each flowchart there are alternative sets of metadata terms that are from selected namespaces for selection. Each chart is followed by the text-based explanations corresponding to the flowchart, with notes and steps in a table, as well as examples whenever necessary. (See an example of the decision trees for creator at: http://aims.fao.org/lode/bd/creator). All decision trees can be found at the Website of FAO AIMS at: http://aims.fao.org/lode/bd.

6. Presenting the recommendations

The recommendations were compiled under a title "LODE-BD Recommendations -- Report on how to select appropriate encoding strategies for producing **L**inked **O**pen **D**ata (LOD)-**e**nabled **b**ibliographic **d**ata" (http://aims.fao.org/lode/bd[4]). The LODE-BD Recommendations aimed to address two questions: how to encode bibliographic data hosted by diverse open repositories for the purpose of exchanging data across data providers; and how to encode these data as LOD-enabled bibliographic data. A selected number of widely used metadata standards and the emerging LOD-enabled vocabularies were used based on the context of the VOA3R community. In the Recommendations version 1.1, metadata terms from the DCMES (dc) and DCMI Metadata Terms (dcterms) are the fundamentals, while metadata terms from other namespaces are supplemented when additional needs are to be satisfied, including the namespaces of bibo (*Bibliographic Ontology)*, ags (FAO's Agricultural Metadata Element set), agls (*AGLS Metadata Standard* of the Australian Government Locator Service), eprint (UKOLN *Eprints Terms)*, and marcrel (*MARC List for Relators)*. All metadata terms used in the LODE-BD Recommendations are presented in a crosswalk table.[5] The report focused on the implementation of standards for data structures (e.g., on which namespace and what properties would best fit for an encoding decision), combined with limited consideration for data contents (e.g., on what metadata properties are mandatory and which value space should consider using controlled vocabulary). It is planned that other reports in the LODE Recommendations series will address other issues, with some being closely related to LODE-BD, such as the need for encoding LOD-enabled authority data and subject vocabularies.

LODE-BD aims to be useable beyond the agriculture and VOA3R communities and plans to include more widely adopted properties from other namespaces, after finishing a study of the usage of the properties in related bibliographic datasets. The next version, LODE-BD 1.2, is under development and will be released within 2011. Meanwhile, the LODE-BD Recommendations report is open for suggestions of new components according to the needs of data providers and of the new development of the LOD community. Currently, VOA3R data providers are consulting the recommendations and preparing their implementation strategies.

This poster will present a portion of the data sheet that aligns various original elements, including the table of the metadata properties and groups, the graphic presentations of the conceptual model, the crosswalk table, and two of the flowcharts. Experiences shared in this poster should provide similar projects with useful tips in methodologies, and may promote more generalized and common LOD-enabled encoding for open bibliographic data.

Acknowledgement:

This work is partially supported by the European Commission through the ICT PSP Grant #250525 for VOA3R (Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment). The authors also would like to thank the support and advice from Ioannis N. Athanasiadis, Nikos Manouselis, Ilias Hatzakis, Tom Baker, Gordon Dunsire, Hugo Besemer, Fernanda Peset, Xavier Agenjo, Francisca Hernández, MacKenzie Smith, Karen Coyle, Antoine Issac, the FAO AIMS Group, and the content providers of the VOA3R team.

[4] LODE-BD Recommendations v.1.1 Report on how to select appropriate encoding strategies for producing **L**inked **O**pen **D**ata (LOD)-**e**nabled **b**ibliographic **d**ata. (2011). Agricultural Information Management Standards (AIMS), Food and Agriculture Organization (FAO) of the United Nations. Available at: http://aims.fao.org/lode/bd

[5] http://aims.fao.org/lode/bd/crosswalk