Applying New Trends to the Management of Bibliographic Information on Agriculture
Die Anwendung neuer Trends im Management von bibliographischen Agrarinformationen

The Food and Agriculture Organization of the United Nations (FAO) is continually working to address the needs of efficient management, discovery and exchange of agricultural information. This paper presents an overview of work undertaken so far using standardized common **metadata** to exchange information in eXtensible Markup Language (XML); and **ontologies** to provide a structured view of domain knowledge that allows for the creation of useful and user-friendly tools for indexing and efficient retrieval of information.

Abstract

With the current rapid developments in the field of information management, it has become important to keep abreast with the new technologies and to apply them appropriately, wherever possible, to improve access to information. This paper gives an outline of the current initiative at FAO that attempts to implement these new technologies with the primary aim of providing improved access to agricultural information. This initiative, which comprises the Agricultural Metadata Element Set (AgMES) and Agricultural Ontology Service (AOS) projects, seeks to provide standards to describe resources in the agricultural domain, by using well-known metadata sets and domain-specific controlled vocabularies, such as thesauri, classification schemes, and ontologies.[1,2] The paper gives a basic overview on the concept of metadata and its importance in facilitating information discovery, and on the use of XML as a platform-independent tool for achieving interoperability between different applications in the domain of agriculture. The second half of the paper introduces ontologies and describes their usefulness in representing and reusing domain knowledge for providing precise concept-based search results.

Bei den gegenwärtigen bedeutenden Fortschritten im Bereich des Informationsmanagements ist es wichtig geworden, den zugehörigen Technologien Rechnung zu tragen und sie, dort wo möglich, sinnvoll anzuwenden, um den Informationszugang zu verbessern. Dieses Papier bietet einen Überblick über die derzeitigen Bemühungen der FAO, diese Technologien zu implementieren, mit dem vorrangigen Ziel,, einen besseren Zugang zu landwirtschaftlicher Information zu schaffen. Diese Initiativen, die aus dem Metadata Element Set (AgMES)- und dem Agricultural Ontology Service (AOS)- Projekt bestehen, versuchen allgemeingültige Standards zur Beschreibung von Resourcen im Bereich der Landwirtschaft durch den Einsatz etablierter Metadatensätze und sonstiger kontrollierter Wörterverzeichnisse wie Thesauri, Klassifikationsschemata und Ontologien bereit zu stellen. [1,2] Der Artikel gibt einen grundlegenden Überblick zum Konzept von Metadaten und ihrer Bedeutung zur Erlangung einer Komparabilität zwischen verschiedenen Anwendungen im Bereich der Landwirtschaft. Der zweite Teil des Artikels stellt Ontologien und ihren Nutzen für die Erzielung von genauen Suchergebnissen vor.

Keywords: FAO, information retrieval, metadata, ontologies, thesauri
FAO, Informationssuche, Metadaten, Ontologien, Thesauri

1. Introduction

The exponential and uncontrolled growth of resources on the World Wide Web (WWW) has brought with it the daunting task of their management. The information needs of users are often complex, implying that information – and eventually knowledge – be drawn from distributed archives and systems and presented to the user in a cohesive and comprehensive manner. The “semantic web” initiative defines this goal, and various projects are under way to improve resource discovery.[3]

Experts in information management and related areas have been working progressively on new strategies to solve the problem arising from information overflow. Our efforts in FAO take into account their successful methodologies and aim to apply them to improve access to information in the domain of agriculture.

There is a great need to create robust mechanisms to exchange and share information stored in departmental, regional, national and international databases. The eXtensible Markup Language (XML) has been gaining widespread acceptance as a platform-independent format for data exchange; eliminating the barriers set up by sometimes authoritarian and proprietary formats.[4] XML uses tags as containers of content or information. Our efforts in the metadata area try to define these human- and machine-understandable tags or containers.

2. Metadata

2.1 What is metadata and why do we need it?

Put simply, metadata is “data about data”. It enables effective, efficient, and accurate use of datasets and data collections, and is the departure point for all discovery systems. In paper-based systems, the card catalogues contain the metadata; in the current digital world, machine-readable repositories contain the metadata. It allows for precise description of resources and the sharing of these descriptions as individual containers, called metadata records, without the necessity of involving the resources themselves in the transaction.[5]

Metadata holds different meanings for different people. To metadata creators, it is a means of moving data from an unorganized pool of resources to an organized pool. Creating metadata helps us to organize knowledge about a given domain and present it at a higher level of abstraction, so that it can be more easily maintained at both the dataset as well as the collection level. To system designers, metadata represents an efficient medium allowing for fast access to resources, especially when resources are “non-searchable” such as images, videos, power point files, audio/sound files. Finally, for users, metadata is a quick way of accessing information without having to go through each resource one by one. Imagine, for example, a library without a card-catalogue with the title, author or subject information about the available books and journals. The only way to find the book that you are interested in would be by looking at every book on the shelf!

2.2 The metadata initiative in FAO

The AgMES project was launched in November 2000, at a workshop in Brussels jointly organized by FAO and OneWorld Europe.[6] The aim of the project is to promote, among other things, the use of metadata through the adoption of standardized agricultural metadata terms, to “simplify resource discovery and interoperability between and among uniquely and richly described agricultural resources”.[7] The project intends to promote the integration of data from different sources, and to engage in – and thus demonstrate the benefits of – effective data exchange. The AgMES metadata defines elements, qualifiers, encoding schemes and controlled lists that are generic, yet necessary, to describe agricultural resources such as projects, presentations, images, best practices, statistical data, maps etc., in all areas relevant to food production, nutrition and rural development. Initial work concentrated on the description of document-like resources, e.g. maps, audio files, images, webpages, books, journals, journal articles, PowerPoint presentations etc.

The primary goal of the AgMES project was to define an interoperability layer, using emerging standards, that facilitates the efficient exchange and use of agricultural content, allowing national agricultural databases and FAO to efficiently exchange information. This means that local databases could be part of the "open exchange of information" scene, without requiring them to change anything in the database itself. This method promotes the mapping of local database structures to a commonly accepted data model and the use of metadata harvesting protocol defined by the Open Archives Initiative (OAI) to harvest this metadata, contained within XML files, for reuse.[8] The OAI promotes heightened "access to e-print archives as a means of increasing the availability of scholarly communication".[9] A simple Dublin Core (DC) format, consisting of the 15 unrefined elements functions as a common denominator to homogenize both harvested raw data or sets of results produced by parallel searching on distributed databases.[10]

The 15 Dublin Core elements and their refinements and schemes were not sufficient to describe agricultural resources. Therefore, we decided to extend this element set with necessary agriculture specific elements, refinements and schemes. Here are brief descriptions of what each of these is:

- Element - refers to a label that describes particular information about a resource, e.g. Title, Creator, Date.
- Refinement – makes the meaning of an element narrower or more precise, e.g. the element subject is made specific by dividing it into subject as defined by a classification scheme and a subject as defined by a thesaurus.
- Scheme - aids in the interpretation of the value of an element, e.g. the use of YYYY-MM-DD as a format for encoding dates of a resource. Schemes are usually controlled vocabularies, controlled lists, and formal notations or parsing rules that allow for further understanding of the content of an element.

The steps involved in defining these new extensions to DC elements included:

- developing a conceptual framework of different types of document-like agricultural information resources to homogenize the description of meta elements;
- evaluating other available standards and common resource description practices to avoid reinvention of the wheel;
- identifying 13 core elements for resource description at generic level and qualification of elements (using the DC guidelines), to enable description of resources at different levels of granularity;
- preparing draft specifications of new elements, refinements and schemes;
- discussing these proposals within the agricultural community, mainly through the AgStandards discussion forum.

AgMES encourages the exchange of metadata in a platform-independent yet machine-readable format (such as XML). Let's look at an example of how content, or parts of a metadata record, can be tagged in XML. Imagine we have a resource called, "*Review of ocean climate research on the role of the Nordic countries in global ocean climate research*". This review article can be described with the following subject keywords: Climatology, Oceanology, and Hydrological Cycle. The first example below shows the "subj" tag being used to encode the above three keywords in XML. The second example uses the tag called "dc: subject" from the DC metadata element set. The tag is further refined with "ags:subjectThesaurus", using the refinements made in the AgMES Project. The tag can be

further qualified with an attribute "scheme" to allow us to specify that these three subject keywords were taken from FAO's multilingual AGROVOC Thesaurus.[11]
Table 1: XML encoding

The goal of the AgMES project is to define intelligently tags, or element names, such as "ags:subjectThesaurus", and propose correct schemes to allow us to know more about the content of these elements.[12] Furthermore, using schemes or controlled vocabularies like AGROVOC helps us to control the *source* of the content and to maintain a level of quality. The elements defined by the AgMES supplement the DC elements. They may be used not only together with DC but also with other metadata standards. This property of metadata allows for easy manipulation and customization of metadata to suit the specific needs of a given application.

Future developments and documentation to support the project will include:
- developing good documentation and user guidelines;
- monitoring the impact of the proposed metadata for agricultural resources, making any changes or improvements based on the results of the impact study;
- registering the new metadata elements in established registries to promote its expansion and uptake;
- launching pilot projects between FAO and agricultural subject gateways to share metadata about agricultural resources in a commonly accepted XML based format;
- creating a metadata clearinghouse for exchange of knowledge and advances on semantic standards in the domain of agriculture.

3. Ontologies

3.1 What are ontologies and why do we need them?

An ontology is a system that contains terms, the definitions of those terms, and the specification of relationships among those terms. It can be thought of as an enhanced thesaurus – it provides all the basic relationships inherent in a thesaurus, as well as defining and enabling the creation of more formal, specific and powerful relationships. An ontology structures the knowledge in a domain and captures the meaning of concepts that are specific to that domain. The application of an ontology into an information management system provides end-users with tools that make indexing and therefore, retrieval, more precise.

3.2 The ontology initiative in FAO

Several satellite ontologies were developed within the broader context of the AOS project at FAO. As there is no standardized procedure yet on how to develop an ontology, different approaches were used in their construction.

Fisheries Ontology: The existence of several vocabularies and terminology tools in the areas of fisheries and aquaculture provided a good basis on which the ontology was built. Using ontologies for the linking and semantic mapping of concepts allows the applications to provide semantically accurate search results.[13]

Food safety ontology: The core ontology, which was built from scratch by subject experts, was extended using a toolset for ontology building. The first tool was a web crawler which identified domain related webpages to define additional relevant terms. A second method to extend the core ontology was to extract keywords from a sophisticated thesaurus on the basis of domain specific documents. To distinguish between domain-specific and non-domain-specific keywords the process was repeated with documents not relevant to the subject.[14]

In addition to traditional relationships, such as equivalence (equivalent terms), hierarchical (broader term/ narrower term) and associative (related term), ontologies also allow for the definition of additional relationships, such as:

- exact equivalence (synonym)
- partial equivalence (words with similar meaning)
- single to multiple equivalence
- inexact equivalence (antonyms)

Defining non-traditional relationships offers the added possibility of distinguishing between concepts that are normally assumed to be similar in meaning. For example, very often "risk analysis" and "risk assessment" are used as synonyms. In a traditional relationship, they may be equated, whereas in a non-traditional or ontological relationship a slight distinction can be made depending on the domain at hand.

In the food safety ontology, the subject experts used the option to define their own relationships, e.g.:

- **compose**: smaller parts making up a piece, element or sub-body of a bigger part
- **follow**: to come before in order, time or position

The expected benefits of taking this approach were to provide support to the end-user in refining their search, and to facilitate navigation within the domain. Furthermore, an ontology editor was used to store the ontology in a database or in Resource Description Framework Specifications (RDFS).[15] The ontology can be accessed through a web interface; this allows the user to browse the ontology hierarchically or to search for a specific term and its relationships.

3.3 Lessons learnt

- **Methodology**

Ontologies emerged several years ago and the involvement of information specialists and the agricultural community is considerable. Over the last few years, great efforts were made to develop tools to facilitate ontology creation and a good theoretical basis. Despite all these efforts there is no well-defined framework available to date. Creating an ontology still involves a trial and error methodology, which may ultimately be the best approach. Additionally, its success also depends to a large extent on available financial and information resources as well as on the chosen domain. Former attempts to build an ontology can only rarely be reused because they are frequently too general and abstract, and therefore not very useful for the creation of concrete ontologies. This lack of methodology and expertise means that the process is invariably slow and can become complicated.

- **Merging of thesauri and ontologies**

The relationships in traditional thesauri (Broader Term, Narrower Term, and Related Term) are defined in the same way and do not cause any particular problems. Risks were commonly encountered when merging thesauri relate to the *location* of the descriptors in the hierarchy.

The FAO and the National Agricultural Library (NAL) thesauri both use the descriptor "animal products".[16] However, the hierarchy related to this term differs in various points which is illustrated in the following table:

Table 2: Comparing the same descriptor in two different thesauri

A particular characteristic of ontologies is the option to create one's own relationships in order to add more information. This can be an advantage, especially in clarifying the link between two terms. However, the freedom of not having to use standardized relationships can become problematic when ontologies are then merged. In these situations, subject specialists have to deal with the problem of merging the relationships correctly.

- **Need for subject experts**

The availability of subject specialists in ontology creation is crucial. Subject specialists have to be experts in the domain in which the particular ontology will be developed because only they are able to decide on the correct terms and relationships among those terms.

Although every subject specialist is an expert in his/her area, the understanding of a term might differ. The experts, therefore, have to pay attention that each of them uses the term under discussion in the same way. Where experts from related, yet slightly different, areas are involved it might even happen that they use the same term but with a completely different meaning depending on the context. Deciding – and agreeing – on relevant terms, their location within the hierarchy and the establishment of the relationships is a very time-consuming task for the experts and should not be underestimated.

4. Conclusion

It is important to understand the meaning behind the axiom "Garbage in = garbage out!" (viz. if invalid data is entered into a system, the resulting output will also be invalid). It is essential to provide not only common metadata elements and share well established thesauri and classifications, but also to make sure the quality is well above par.

Metadata allows users to retrieve not only the resources related to the search query but also to obtain adequate information on the content of the resource, such as the date of publication, name of the author(s) or languages in which the content of the resource are available. This information is of even of greater importance when the actual resource is not available online but resides physically at a different location.

Collaborative efforts are needed to describe different types of resources. Metadata elements which are useful for the description of books are not necessarily the best for describing a shipping vessel in a fisheries' database. The AgMES project is challenging, but it will only be successful if conscientiously and continually maintained, applied and updated. Under the AgMES umbrella, other projects will be able to create new terms for other knowledge areas and for other types of information resources within the domain of agriculture. They will be represented and documented as they become available for reuse.

Ontologies open up new perspectives in information management because they have the potential to add new semantic-based features to an information system. However, the ontology is a new methodology and still in its infancy; common guidelines on ontology creation are not yet well established. Information specialists will have to combine their knowledge in information management with the domain specific expertise of subject specialists, and work closely together to achieve functional and successful results.

Acknowledgements

The authors wish to acknowledge the assistance of Fynvola Le Hunte Ward in the preparation of this article, as well as the support of all the members of the GILW, AGRIS/CARIS and Documentation Unit.

References:

[1] Agricultural Metadata Element Set: http://www.fao.org/agris/agmes/
[2] Agricultural Ontology Service: http://www.fao.org/agris/aos/
[3] The Semantic Web initiative: http://www.semanticweb.org/
[4] XML Specifications from W3C: http://www.w3.org/XML/
[5] NSDL metadata primer: http://metamanagement.comm.nsdlib.org/overview2.html
[6] One World Europe: http://www.oneworld.org/europe/en/
[7] AgMES Concept Note: ftp://ext-ftp.fao.org/agris/agmes/AGMESConceptNote.pdf
[8] Open Archives Initiative: http://www.openarchives.org/
[9] OAI and OAIS: What's in a Name?, Peter Hirtle http://www.dlib.org/dlib/april01/04editorial.html
[10] Dublin Core Element Set: http://www.dublincore.org/documents/dces/
[11] AGROVOC Multilingual Thesaurus: http://www.fao.org/agrovoc/
[12] AgMES elements: http://www.fao.org/agris/agmes/Documents/AgMESv1.1.htm
[13] Fishery Ontology Project: http://www.fao.org/agris/aos/Presentations/Fishery_Status3.doc
[14] Creating an ontology on Food Safety, Animal and Plant Health (OFsAPH): http://www.fao.org/agris/aos/Presentations/dc2002.ppt
[15] Resource Description Framework Specifications (RDFS): http://www.w3.org/RDF/
[16] NAL Thesaurus http://agclass.nal.usda.gov/agt/agt.htm

Authors:

Ms. Tanja Wildemann
Associate Professional Officer, Food Quality/Safety & Information Management
Library and Documentation Systems Division (GIL)
Food and Agriculture Organization of the United Nations
Viale delle Terme di Caracalla
00100 Rome, Italy
Fax: +39 06 570 54593
Tel: +39 06 570 53704
Email: Tanja.Wildemann@fao.org

Ms. Gauri Salokhe
Information Management Consultant
Library and Documentation Systems Division (GIL)
Food and Agriculture Organization of the United Nations
Viale delle Terme di Caracalla
00100 Rome, Italy
Fax: +39 06 570 54049
Tel: +39 06 570 54183
Email: Gauri.Salokhe@fao.org

Dr. Johannes Keizer
Information Systems Officer
Library and Documentation Systems Division (GIL)
Food and Agriculture Organization of the United Nations
Viale delle Terme di Caracalla
00100 Rome, Italy

Fax: +39 06 570 54049
Tel: +39 06 570 52729
Email: Johannes.Keizer@fao.org

Table 1: XML encoding

Ex. 1	<subj> Climatology <subj> <subj> Oceanology <subj> <subj> Hydrological Cycle <subj>
Ex. 2	<dc:subject> <ags:subjectThesaurus scheme="ags:AGROVOC"> Climatology </ags:subjectThesaurus> <ags:subjectThesaurus scheme="ags:AGROVOC"> Oceanology </ags:subjectThesaurus> <ags:subjectThesaurus scheme="ags:AGROVOC"> Hydrological Cycle </ags:subjectThesaurus> </dc:subject>

Table 2: Comparing the same descriptor in two different thesauri

AGROVOC (FAO)	NAL Agricultural Thesaurus (USA)
Animal products	**Animal products**
Uf: livestock products	*BT: animal products, byproducts and wastes, agricultural products*
NT1: eggs	*NT: animal fats and oils*
NT1: fish	*NT: animal protein concentrates*
NT1: hive products	*NT: animal-based foods*
NT: ivory	*NT: fish products*
NT1: meat	*NT: non-food animal products*
NT2: beef	
NT2: buffalo meat	
RT: agricultural products	*RT: animals*
RT: bioproducts	
RT: charcoal	
RT: milk	
RT: processed animal products	
RT: products	