Ontology-based Navigation of Bibliographic Metadata: Example from the Food, Nutrition and Agriculture Journal

Margherita Sini[1], Gauri Salokhe[1], Christopher Pardy[1], Janice Albert[1], Johannes Keizer[1], Stephen Katz[1]

[1] Food and Agriculture Organization of the United Nations, Rome, Italy
{margherita.sini, gauri.salokhe, christopher.pardy, janice.albert, johannes.keizer, stephen.katz}@fao.org

Abstract. This paper describes the work done within the Food and Agriculture Organization of the United Nations (FAO) on providing an ontology-based navigation for the Food, Nutrition and Agriculture (FNA) Journal. The aim of the revised navigation was to provide more efficient and effective browsing of the Food and Nutrition Publications using a knowledge model to guide the user with concepts and relationships relevant to a specific subject area. With this approach, data from two different bibliographical databases was merged, unified and presented to the user with improved services. A preliminary metadata merge was needed to combine all the information into one system in order to produce a metadata-ontology. Resource Description Framework Schema (RDFS) was chosen to exploit semantic relationships, e.g. the possibilities of browsing the data in different ways (by keywords, categories, authors, etc.), and the creation of a multilingual concept-based advanced search.

Keywords: Ontology, Semantic Browsing, Semantic Search, RDFS.

1 Introduction

A hypertext-based system originally created to facilitate access and exchange of scientific data, the World Wide Web (WWW) has become a main communication and information resource for use by the general public. The exponential growth in the available resources on the Web makes it necessary to extend the capability of computers in order for them to understand the information better and serve up the best results to the users. However, this capability of computers to process information on the Web is limited by:

- ambiguity in the meaning of search strings: for example, a search for ‘rice’ in Google, returns results which include, inter alia, a surname, a university, a research institute, or a crop;

- a high recall without logical order, for example, Google returns over 116 000 000 results for 'rice' without organizing the results by type or category;
- the inability to understand the user's intended meaning;
- no guarantee of the trustworthiness of information supplied; and
- lack of assistance in the formulation of better queries.

Thus, formalizing the semantics or meaning of data so that it is readable by computers improves not only the way in which information is organized and displayed but also how it is processed. Computers need to be provided with explicit context for terms, such as their attributes and their relations to related terms. Making explicit the properties and relations of terms provides a starting point for their conversion into ontology.

1.1 Problem background

The articles in the Food, Nutrition and Agriculture (FNA) Journal cover topics such as community nutrition, food quality and safety, nutrition assessment, nutrient requirements, food security and rural development. The full-text articles may be in English, French or Spanish. The Food and Agriculture Organization of the United Nations (FAO) has provided information about nutrition and food safety to audiences around the world for over four decades. A simple newsletter initiated in the 1960s evolved into a multilingual, interdisciplinary journal during the 1980s and 1990s. By the year 2004, the FNA Journal publishes articles on diverse topics such as nutrient requirements, nutrition education, food safety and international food standards. Today, over 6 000 universities, research institutions, government agencies, non-governmental organizations and food companies subscribe to FNA Journal. Initially, the journal was available only in print, but since 1997 is journal is available on FAO's Web site. And recently, some semantics was incorporated into its navigation facility.

1.2 Benefits of ontology usage

Ontology is a model of organized knowledge in a restricted domain (e.g. fisheries, nutrition, and medicine). Ontologies consist of components such as "concepts", "attributes", "relations" and "instances". In an ontology, concepts correspond to objects to be organized (e.g. projects, people, products, etc.); attributes are the traits of those objects (e.g. titles, addresses, colour, age, etc.); relations connect two objects or an object and a property to each other (e.g. « Person » can be linked through an « employed by » property to an «

Organization »); and instances are the actual data in an information system (e.g. “John Smith”, “FAO”).

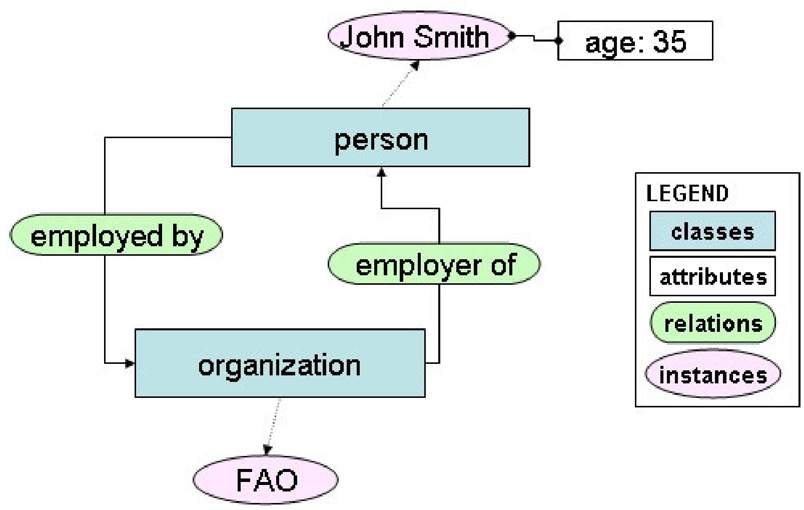

Figure 1. Example of ontologically-expressed relationships between persons and organizations

Ontologies increase the efficiency and consistency of describing resources, by enabling more sophisticated functionalities in development of knowledge management and information retrieval applications. The use of standards, such as Resource Description Framework (RDF Primer, 2004), RDF Schema (RDF Vocabulary, 2004) and Web Ontology Language (OWL, 2004), provide structures for sharing common descriptions, definitions and relations within the agricultural community. For the context of FNA, the RDFS format has been chosen as it is best suited for its needs and purpose.

2. Agricultural information systems and common standards

FAO has a normative role to play in the creation and promotion of standards in agricultural information management. It understands itself as an organization responsible for knowledge brokering. In 1995, the Online Computer Library Center organized the first workshop of the Dublin Core community with the aim of finding a core set of metadata elements to describe networked resources (DC1, 1995). This discussion was taken up in a meeting on agricultural standards, organised by Oneworld Europe in collaboration with FAO in Brussels in 2000 (Agstandards Workshop, 2000). The meeting raised awareness amongst information providers of the new opportunities for sharing information through use of metadata standards and platform-independent formats such as XML with the agriculture community in mind. The result of this is available in the Agricultural Information Management Standards Web site (AIMS, 2005). Within the framework of AIMS, the idea of implementing semantic standards-based navigation of FAO's information sources is now a priority.

2.1 Background study and the bibliographic model

In 1998, Weinstein generated a knowledge base of metadata from a sample of Machine Readable Cataloguing (MARC) records (Weinstien, 1998). He implemented the ontology in description logic, a knowledge representation language, and mapped MARC attributes and values to the ontology. In this way, ontology was adapted to describe metadata.

ABC ontology and its model were developed within the Harmony international digital library project in Cornell University. It was designed to offer a common model that would facilitate the ability to exchange metadata ontologies from different domains (Lagoze & Hunter, 2001). One of the results of the project was a metadata model with more logically described time and entity semantics. Based on this model, a metadata repository of RDF descriptions was built with a search interface on top.

One of the first metadata applications in the domain of Agriculture, based on Dublin Core, was the AGRIS application profile (AGRIS AP, 2005). The AGRIS Application Profile is an XML-based bibliographic metadata exchange format that allows sharing of information across dispersed bibliographic systems. This is a major step towards exchanging high-quality and medium-complexity metadata in an application independent format and provides possibilities to offer value-added services, irrespective of how the information was stored locally. The AGRIS Application Profile was used as model for the underlying bibliographic ontology design.

2.2 Bibliographic metadata about the FNA Journal articles

Metadata about each article in the FNA Journal was available in FAO's Corporate Document Repository (FAO, 2006a) and the FAO's bibliographical database (FAO, 2006b). The two metadata sets contained slightly overlapping metadata information about each resource. The original formats of each database were proprietary; XML, in the case of Corporate Document Repository and tag-text in the case of FAO Bibliography. The goal was to combine metadata from these two bibliographical databases and convert it to a single RDFS format. As many of the articles (and their metadata) in the journal are available in English, French and Spanish, we used UTF-8 for ontology storage. The best option to store and manage the bibliographic multilingual data in RDF was the Karlsruhe ontology (KAON) suite of tools. Although at that time Protégé was more compliant to strict RDF, it did not allow for management of multilingual data and was thus not used.

2.3 KAON Suite: An Ontology modeling infrastructure

The modelling infrastructure of KAON which supported an enhanced version of RDF ontologies was developed by the University of Karlsruhe (Germany) in 2002 (KAON, 2004). This project used the graphical ontology editor OI-Modeler and, at that time, the tool produced proprietary KAON specific RDF metadata.

3. Making the semantics explicit

As mentioned earlier, the FNA journal when first posted on FAO's website in 1997, did not provide any semantics-based navigation to its users. To allow its on-line readers to discover information embodied in the articles, it was necessary to make explicit the relationships between meta-data about the articles. An ontology was developed using available article-level meta-data from FAO's cataloguing and indexing systems. The ontology was originally modelled using Topic Maps, but subsequently implemented using RDF (XTM, 2001).

The AGRIS Application Profile provided the backbone on which agricultural resources, namely the individual articles from the FNA Journal, could be organized and described. A "metadata ontology" using concepts from the AGRIS Application Profile was created (see Figure 2). The ontology is composed of: concepts, relationships between concepts and instances, which were the actual metadata records described in section 2.2.

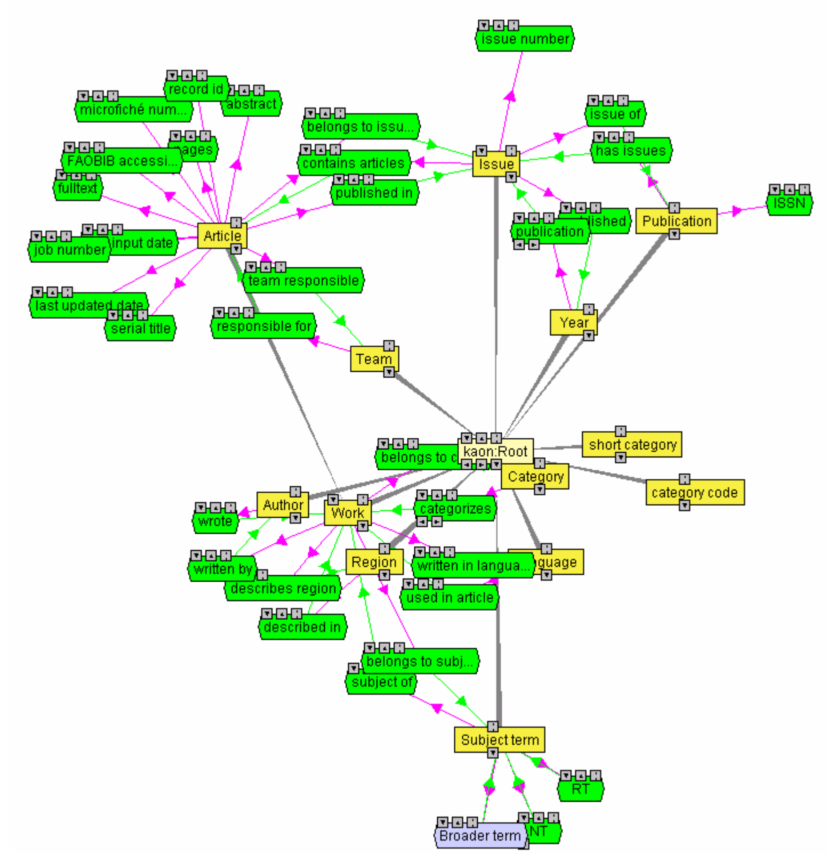

Figure 2. Part of the FNA ontology describing a bibliographic record

The ontology consists of all elements or concepts contained in the AGRIS Application Profile necessary to describe a bibliographic record. Following the creation of this "meta-ontology", instances were created taking the actual metadata merged from the FAO's Corporate Document Repository and the FAO Bibliographical Database.

```
<kaon:Documentation rdf:ID="i-1079447777762-1153656351"
kaon:value="6598FR.txt">
   <kaon:inLanguage rdf:resource="&kaon;fr"/>
   <kaon:references rdf:resource="#id6598"/>
</kaon:Documentation>
<kaon:Documentation rdf:ID="i-1079447777762-1153656351"
kaon:value="6598EN.txt">
```

```
   <kaon:inLanguage rdf:resource="&kaon;en"/>
   <kaon:references rdf:resource="#id6598"/>
</kaon:Documentation>
<kaon:Documentation rdf:ID="i-1079447801627-1878849725"
kaon:value="6598ES.txt">
   <kaon:inLanguage rdf:resource="&kaon;es"/>
   <kaon:references rdf:resource="#id6598"/>
</kaon:Documentation>
```

Figure 3. Section of the ontology describing languages in which an article is available (KAON specific RDFS)

The steps taken to merge the metadata records from two databases are as follows:

1. An export was made from each database in a flat "tagged text" format.
2. Then, the records were merged from both files using the FAO Job Number, a unique number which identifies the documents.
3. In the event of conflicting metadata, e.g. in two titles, the information from FAO Bibliographical Database was taken as the metadata created by trained cataloguers.
4. In the event where there is no conflict, the metadata from both records was retained.
5. XML Transformation was used to convert the metadata records into the records for RDFS model.

A portal was created to implement search and browse functionalities for the FNA Journal on top of the RDFS metadata (FNA Journal, 2005).

3.1 Searching the system

A search application was created on top of the "meta-ontology" and the "instance" data in RDFS. When a user enters a phrase in the query box, the system searches all the objects in the ontology which may be lexicalized with the query string. The objects can be articles, publications, authors, keywords, subject categories, languages or geographical areas. When one or more matches are found, the user is provided with the list of objects.

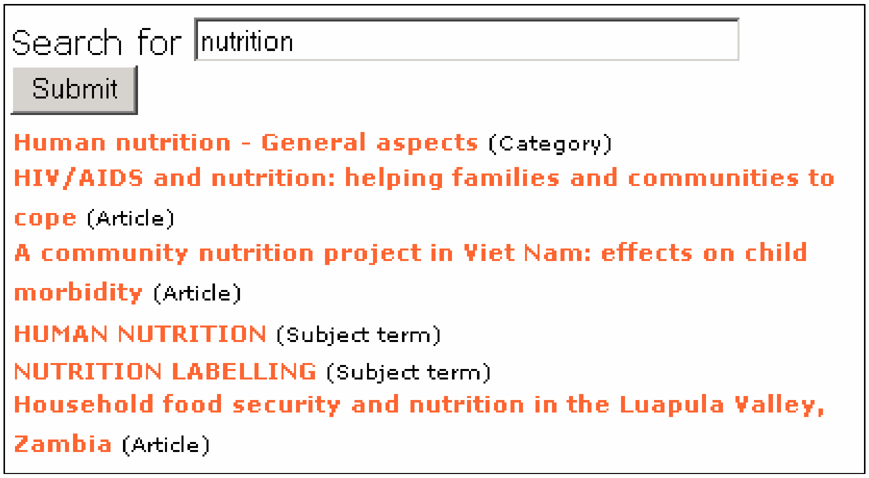

Figure 4. Example of multilingual concept resolution

3.2 Browsing the system

The portal is predominantly "browse-based" although users can also search the metadata using a free-text search. A user is guided through the navigation of data by following the links that connect the different metadata elements, such as the articles published within a specific issue, the authors of an article or its available languages, all articles associated to a specific keyword, etc. The user, during the navigation process, is able to get answers to her questions such as "give me all articles published in this issue", "give me all authors that wrote about this article", "list all articles about 'allergens'", etc. with just a single 'click'. The example below depicts how a user can find "all articles about CONTAMINATION" by clicking on the keyword.

Figure 5. Example of single click queries

Although these functionalities may also be implemented using a traditional relational database approach, the semantic backbone allows for implementation of additional functionalities (see section 3.3), and from a technical point of view implementation is made easier. For example, by just updating the ontology, with more instances, properties or relationships, the system automatically displays the new “connections” with no further technical intervention on the dynamic web pages.

One major advantage of an ontology-based system over a traditional system is in its ability to perform concept-based searches. The FNA portal allows searches for specific concepts; for example, a search for “child” or “children” will give same result even though these are two different lexicalizations of the same concept, that is one is the plural of the other. The portal also allows multilingual searches: “children” or “enfants” will retrieve the same result.

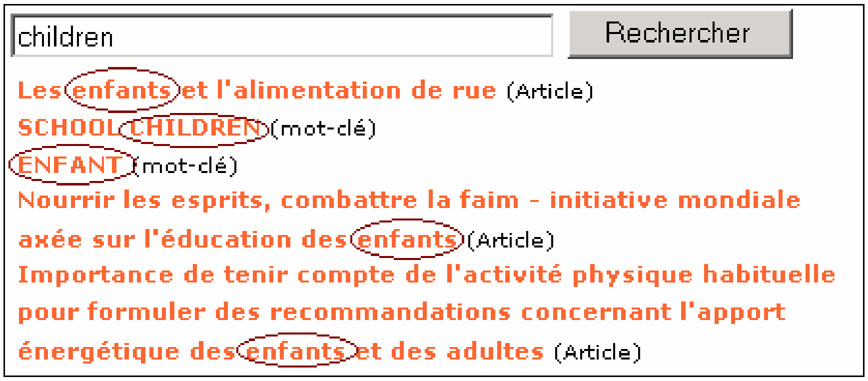

Figure 6. Example of singular vs. plural concept resolution across languages

Once the concept in which the user is interested is identified, he will be directed to articles that are relevant to that concept even if the document may be indexed with keywords in different forms or languages.

3.2 Technical implementation

As mentioned before, the backbone of the system is an ontology written in RDFS. Concepts, relations, attributes, and lexicalizations were accessed programmatically using the KAON Application programming interface developed at the University of Karlsruhe. Specific queries to the system are done using the KAON query language (KAON, 2004). The dynamic pages of the interface were developed using the JavaServer Pages (JSP) programming language (JavaServer Pages, 2006).

In the FNA Journal ontology, all concepts are represented as classes, and the metadata values are represented with instances of those classes. Technically, this meant that we could use a single functionality regardless of the object navigated (e.g., author, issue, articles, and keywords). Therefore, the implementation and deployment process for this portal and required effort. All relationships are also included in the ontology with labels in multiple languages, so that not only the content of the metadata appear in the language chosen by the user while navigating the portal in its preferred interface language but also all the tags identifying the elements are in the languages of users choice (e.g. “Language” in English, “Langue” in French, and “Idioma” in Spanish interface).

3.3 Further functionalities

The use of ontologically organized "bibliographic" metadata about each journal issue and the articles within provided for the following functionalities:

- Easy navigation of the journal issues by following the semantic links;
- Display of articles indexed with the same set of keywords;
- Resolution of user's query terms to a controlled vocabulary (namely the AGROVOC Thesaurus);
- Simple natural language processing as part of the understanding process of the user's query (e.g., distinguishing between the substantive and the function words);
- Cross-linguistic information retrieval;
- Display of semantically related concepts;
- Guided query formulation using the relationships in the ontology; and
- Inferencing, e.g. the user can get the authors associated with specific keywords or vice versa, co-authors can be inferred knowing only the article name and one author.

4. Future work

In a demo version of the portal, which was presented at the ECDL conference in 2004, but which has not yet been published online, other functionalities were available to the users (Kaloyanova, Sini, & Keizer, 2004). For instance, the ability to provide semantically related concepts while navigating the keywords, the ability to provide co-authors, the ability to create a query using graphical-composer (see Figure 5 below).

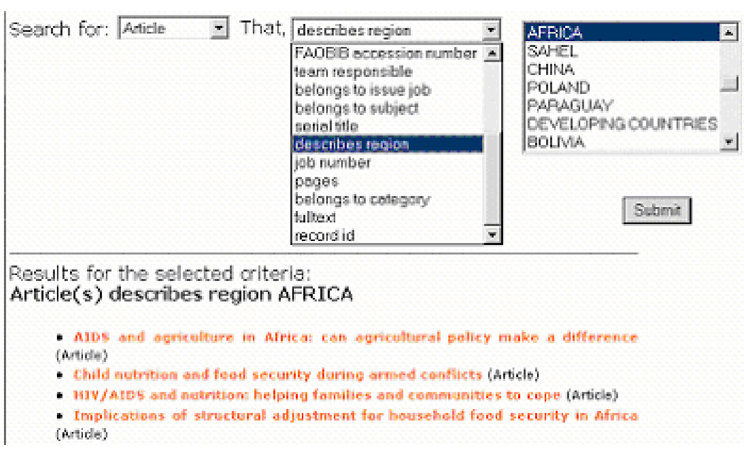

Figure 5 Graphical query composer based on the content of the ontology

The example in Figure 5 illustrates the basic triple-based architecture [predicate, subject, object] of the RDF language. In the above example, this can be noted as [describes region, article, Africa] and results in a list of all "articles which describe the region of Africa". Similar searches can be conducted on other objects like publications, authors, keywords, subject categories, languages or geographical areas.

These steps are just a starting point for further exploitation of other semantic relationships available in a bibliographic metadata record. Making use of existing semantic relationships between, for example, author and keyword, that are not normally exploited in bibliographic databases allowed for more meaningful and hence user-friendly browse experiences. The possible benefits of converting from RDFS to OWL are currently being explored.

References

[1] RDF Primer (2004). Retrieved January 23, 2006, from http://www.w3.org/RDF/

[2] RDF Vocabulary Description Language 1.0: RDF Schema (2004). Retrieved January 23, 2006, from http://www.w3.org/TR/rdf-schema/

[3] OWL Web Ontology Language (2004) Retrieved January 23, 2006, from http://www.w3.org/TR/owl-features/

[4] *DC1: The Essential Elements of Network Object Description* (1995) OCLC/NCSA Metadata Workshop Retrieved January 23, 2006, from http://dublincore.org/workshops/dc1/

[5] *Agstandards Workshop*. (2000). Information Management Standards for Electronic Publishing in Agriculture. Retrieved January 23, 2007, from http://www.idmlinitiative.org/docs/agstandards.doc

[6] *AIMS: Agricultural Information Management Standards Web site*. (2005). Retrieved January 23, 2007, from Food and Agriculture Organization of the United Nations Web site: http://www.fao.org/aims/

[7] Weinstein, P. C. (1998). Ontology-Based Metadata: Transforming the MARC Legacy. *Proceedings of the Third ACM International Conference on Digital Libraries*, 254 - 263. Retrieved January 23, 2007, from Citeseer Web site http://citeseer.ist.psu.edu/212498.html

[8] Lagoze, C., & Hunter, J. (2001). The ABC Ontology and Model. *Journal of Digital Information*, 2(2), article no. 77. Retrieved January 23, 2007, from http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/lagoze-final.pdf

[9] AGRIS AP: The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology Guidelines on Best Practices for Information Object Description

(2005). Retrieved January 23, 2006, from http://www.fao.org/docrep/008/ae909e/ae909e00.htm

[10] FAO (2006a) Corporate Document Repository. Rome, Italy. Retrieved January 23, 2006, from http://www.fao.org/documents/

[11] FAO (2006b) Bibliographical Database. Rome, Italy. Retrieved January 23, 2006, from http://www4.fao.org/faobib/

[12] KAON (2004). KAON Suite. Retrieved January 23, 2007, http://kaon.semanticweb.org/

[13] XTM (2001). XML Topic Maps 1.0. Retrieved January 23, 2007, from http://topicmaps.org/xtm/

[14] *Food Nutrition and Agriculture Journal (FNA Journal).* (2005). Retrieved January 23, 2007, from Nutrition and Consumer Protection Division of FAO Web site: http://www.fao.org/ag/agn/publications/fna/

[15] *JavaServer Pages Technology*. (2006). Retrieved January 23, 2007, from Sun Developer Network (SDN) Web site: http://java.sun.com/products/jsp/

[16] Kaloyanova, S., Sini, M., Keizer, J. (2004) Using RDF to get more from existing bibliographic metadata. *Poster at the European Conference on Digital Libraries*. Retrieved January 23, 2007, from http://www.ecdl2004.org/posters-demos.html