Ontology-based navigation of bibliographic metadata: example of 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 to 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 were reused, by merging and unifying them and make them better accessible to users. 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) has been 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: Metadata, Ontology, RDFS, semantic browsing, Semantic search

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, when searching for ‘rice’ in Google, the user sees results where ‘rice’ is returned, inter alia, a surname, a university, a research institute, or a crop;

- high recall without logical order , for example, Google returns 13,900,00 results for ‘rice’ without organizing the results by type or category;
- 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 surrounding ones. Making clear the properties and relations of terms provides a starting point for their conversion into ontology.

1.1 Ontology definition

Ontology is a model of organized knowledge in a given 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”).

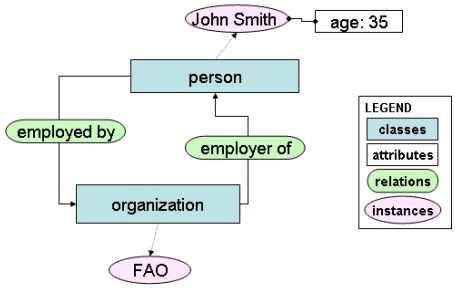

Figure 1 Example of ontologically expressed relationships between persons and organizations

Ontologies help to increase the efficiency and consistency of describing resources, enabling more sophisticated functionalities in development of knowledge management and information retrieval applications. The use of standards, such as Resource Description Framework (RDF) [1], RDF Schema (RDFS) [2] and Web Ontology Language (OWL) [3], provide structures for

sharing common descriptions, definitions and relations within the agricultural community. In the FNA application, the RDFS format has been chosen as it best suited the needs and purpose of the application.

1.2 The Food, Nutrition and Agriculture Journal

The Food, Nutrition and Agriculture Journal (hereafter referred to as FNA Journal) covers topics such as community nutrition, food quality and safety, nutrition assessment, nutrient requirements, food security and rural development. All the articles in the Journal appear as full-text in English, French or Spanish. FAO has provided information about nutrition and food safety to audiences around the world for over four decades. A simple newsletter that was initiated in the 1960s evolved into a multilingual, interdisciplinary journal during the 1980s and 1990s. By the year 2004, the FNA Journal contained articles on topics as diverse as nutrient requirements, nutrition education, food safety and international food standards. Over 6000 universities, research institutions, government agencies, non-governmental organizations and food companies were subscribers to Food, Nutrition and Agriculture. Initially, the journal was available only in print. FAO began to post the journal on its website in 1997 but no semantics-based navigation was available to the users.

2. Agricultural information systems and common exchange standards

FAO has 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 [4] organized the first workshop of the Dublin Core community with the aim of reaching a consensus on a core set of metadata elements to describe networked resources. This discussion was taken up in a meeting on agricultural standards, organised by Oneworld Europe [5] in collaboration with FAO in Brussels, during autumn of 2000 [6]. 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.

Since the Brussels meeting in 2000, FAO has undertaken a number of initiatives to facilitate the standard setting process. More recently, in October 2005, a meeting was organized wherein key players in the area of agricultural information management discussed the implications of the dispersed

developments in exchange standards, tools and methodologies, thus introducing the notion of "coherence". It was agreed that to achieve this coherence, it was necessary to facilitate collaboration, partnership and networking among partners by promoting information exchange and knowledge sharing and harmonize the decentralized efforts currently taking place in the development of methodologies, standards and applications for management of agricultural information systems; consequently, providing a 'one-stop' access to system designers and implementers. The result of this is available in form of the "Agricultural Information Management Standards" Website [7].

2.1 The AGRIS Application Profile

One of the first metadata applications, based on Dublin Core, was the AGRIS application profile [8]. The AGRIS Application Profile is an XML-based bibliographic metadata exchange format that allows sharing of information across dispersed bibliographic systems. It 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.

2.2 Bibliographic metadata about the FNA Journal articles

Metadata about each article in the FNA Journal was available in FAOs Corporate Document Repository [9] and the FAO Bibliographical Database [10]. 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 Unicode for ontology storage.

2.3 KAON Suite

KAON [11] stands for the Karlsruhe ontology, an ontology modelling infrastructure developed by the University of Karlsruhe (Germany) [12]. It was first developed in 2002 and supported an enhanced version of RDF ontologies. 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 stated before, FAO began to post the journal on its website in 1997 but no semantics-based navigation was available to the users. In order to facilitate the discovery of information embodied in the articles by on-line readers, a solution was sought to allow expression of the relationship among 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 [13], but subsequently implemented using RDF as, at the time of the project, Topic Maps did not have a conceptual model or a query language.

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" taking 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.

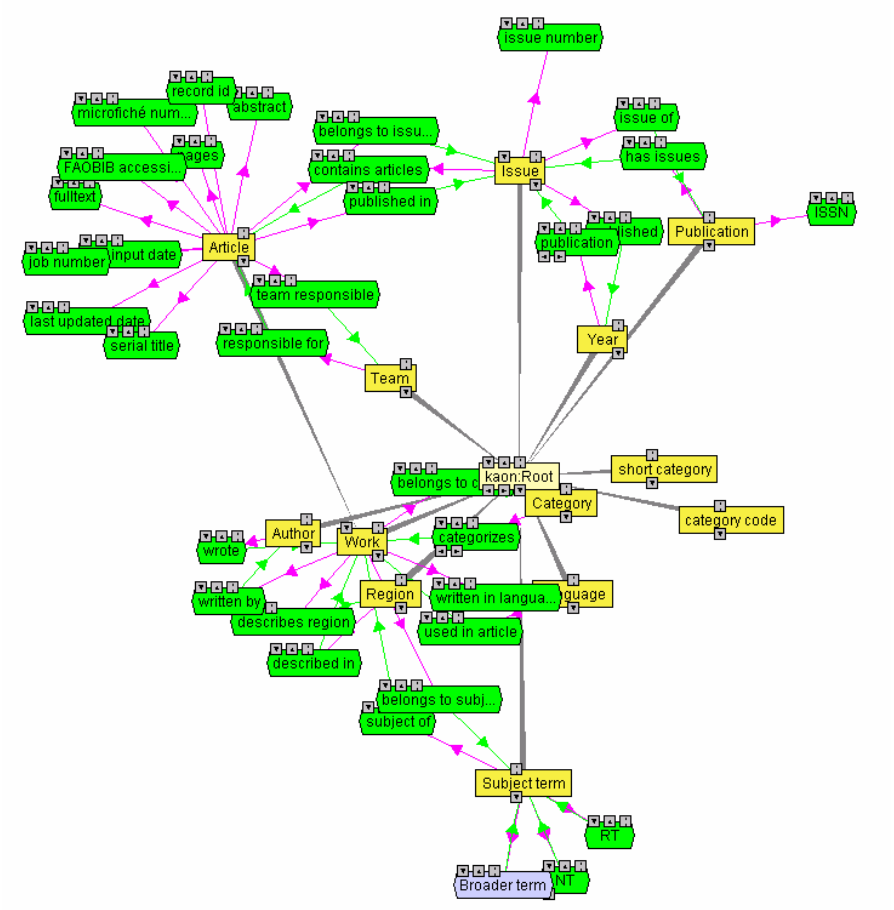

Figure 2. Section of the ontology describing a bibliographic record

The ontology consists of all the elements or concepts contained in the AGRIS Application Profile and necessary to describe a bibliographic record. Following the creation of this “meta-ontology”, instances were created taking the actual metadata merged from the FAOs 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="6598FR.txt">
```

```
   <kaon:inLanguage rdf:resource="&kaon;fr"/>
   <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)

A search application was then created on top of the "meta-ontology" and the "instance" data in RDFS.

3.1 Searching the system

A portal [14] was created to implement search and browse functionalities for the FNA journal. The portal is predominantly "browse-based" although users can also search the metadata using keywords. 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.

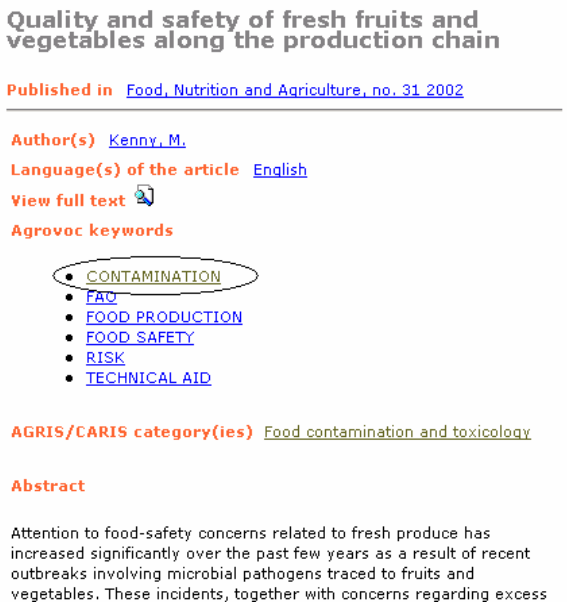

Figure 4. Example of single click queries

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 makes implementation much more easy. For instance, 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 of the real advantages of an ontology-based system over a traditional system is its ability to perform concept-based searches. The FNA portal allows searching for a specific concept; for example, searching for "child" or "children" will give same result even though these are two different lexicalizations of the same concept, although one is the plural of the other. The portal also allows multilingual searches: "hypersensitivity" or "reacciones alergicas" will retrieve the same result. Once the concept in which the user is interested is identified, she will be pointed to the corresponding articles associated to that concept even if the document might 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 [15] developed at the University of Karlsruhe. Specific queries to the system are done using the KAON query language [11]. The dynamic pages of the interface were developed using the JavaServer Pages (JSP) programming language [16].

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 has also been represented 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 unimportant terms);
- 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, that has not yet been published online, which was presented at the ECDL conference in 2004 [17], other functionalities were available to the users. 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).

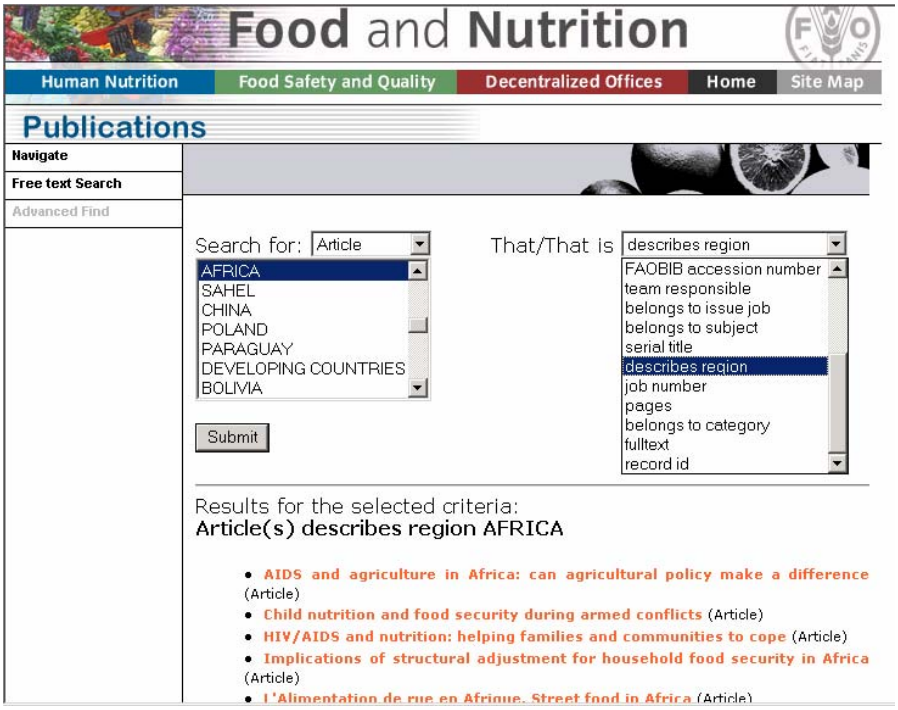

Figure 5 Graphical query composer based on the content of the ontology

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] Resource Description Framework (RDF) http://www.w3.org/RDF/

[2] RDF Vocabulary Description Language 1.0: RDF Schema http://www.w3.org/TR/rdf-schema/

[3] OWL Web Ontology Language http://www.w3.org/TR/owl-features/

[4] DC1: OCLC/NCSA Metadata Workshop: The Essential Elements of Network Object Description http://dublincore.org/workshops/dc1/

[5] Oneworld Web site: http://www.oneworld.net/

[6] Information Management Standards for Electronic Publishing in Agriculture (Practical Workshop): http://www.idmlinitiative.org/docs/agstandards.doc.

[7] The Agricultural Information Management Standards Web site: http://www.fao.org/aims/

[8] The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology Guidelines (July 2005) http://www.fao.org/docrep/008/ae909e/ae909e00.htm and http://www.fao.org/agris/tools/AGRIS_AP/WhatItIs.htm.

[9] The FAO Corporate Document Repository http://www.fao.org/documents/

[10] FAO Bibliographical Database http://www4.fao.org/faobib/

[11] KAON Suite http://kaon.semanticweb.org/

[12] University of Karlsruhe, Germany. http://www.uni-karlsruhe.de/index_en.php

[13] Topic Maps http://topicmaps.org/xtm/

[14] The Food Nutrition and Agriculture Journal: http://www.fao.org/ag/agn/publications/fna/

[15] KAON API http://nlp.stanford.edu/nlp/javadoc/kaon-docs/

[16] JavaServer Pages Technology http://java.sun.com/products/jsp/

[17] Stefka Kaloyanova, Margherita Sini, Johannes Keizer, "Using RDF to get more from existing bibliographic metadata", Poster at the ECDL conference, 2004, http://www.ecdl2004.org/posters-demos.html