TOWARDS A COMMON ONTOLOGY IN AGRICULTURAL DOMAIN: MERGING PRODUCTIVE CHAIN ONTOLOGIES

Autor: Kleber Xavier Sampaio de Souza
Empresa: Embrapa Informática Agropecuária.

Abstract

This article will discuss the usage of the Information Agency as a space for exchange of knowledge among farmers, researchers and rural extension technicians. It will suggest how the information contained in knowledge trees can be improved to become true ontologies in the sense of the Semantic Web Initiative, and present common points identified in the knowledge trees already specified for different domains. The starting point is the Information Agency Project in execution by the Brazilian Agricultural Research Corporation – Embrapa. Although that project has received positive feedback, it would benefit largely if its organization had considered thoroughly the concepts proposed in the Semantic Web initiative. Ontologies for agricultural domain are almost inexistent. Their achievement and the solution of merge related problems are crucial to the construction of more intelligent systems. In this particular case, the Agency will be able to interconnect elements in an ontology and infer that if brachiaria decumbens is a kind of foraging plant for beef cattle, and that foraging is an animal nutrition alternative, than brachiaria is one alternative for cattle.

Keywords

Semantic Web, Ontology, Hyperbolic Tree Visualization, Dublin-Core Metadata Standard

Background

The success of the Web as an attractive environment can be credited mainly to the implementation of the idea launched for the first time by Vanevar Bush in 1945 (Rada, 1991), called hypertext. Working by associating objects instead of simply classifying them hierarchically, hypertext operates closely to the way we think. Naturally, Bush's hypertexts were not associated with today's nor to Internet for they did not exist at that time, but rather to set of texts from several people which were interconnected forming what he called macrotext. Ted Nelson mentioned the expression hypertext for the first time only in 1967.

However, when one carries out a search in well known information portals, the results show low rates of revocation and precision. The quality improvement of the retrieved information is not an easy task, mainly because relevance is measured by the proximity between the concept searched, which is an idea, and the information stored in the system, which is composed of symbols. Computers are symbolic processing machines and, as such, users interact with systems that operate in a level where the relation between an object and the sign (Peirce, 2000) used to represent it is purely arbitrary. That fact is pointed out by Peirce as a *semiotic cut* (Tenorio, 1998).

Although the Web was designed as an information space for both human and computer consumption (Berners-Lee, 1998), most information is structured only for humans. Pages use markup to define structure and presentation of documents. They do not contain, in a machine understandable format, information like subject, keywords or relations among most relevant terms. The objective of the Semantic Web initiative is to create languages that will enable the expression of information semantics by representing associations among elements that could at first sight not seem related.
Even when generated from relational databases and formatted in html pages in the form of tables, the underlying information regarding the relations among the elements is lost, i.e. it cannot be inferred by the machine.

Natural Language Processing research programs have made significant progress, but still do not provide enough material for the construction of web robots that could search the Web and build their own indexes in accordance to someone's preferences (SHOE, 2003). Even if they could

process intelligently textual information, there remains the problem of the information embedded in pictures, graphics and figures, which are out of that research program.

The Brazilian Agricultural Research Corporation - Embrapa - is the largest agricultural research corporation in Latin America and one of the largest in the world. It has forty research units distributed throughout the country connected via satellite network and more than two thousand research scientists working on different areas, products and ecosystems, connecting agriculture, livestock farming, agroindustry and environment.

Embrapa possesses a huge information stock produced along its more than thirty years of existence. That fact, associated to the strategic importance agribusiness represents to Brazil, which plays an important role not only in trade balance but also as a regional developing factor, imposes an enormous challenge to the corporation: to guarantee that knowledge be transformed into information products adequate to farmers, contributing to sustainable development.

The Information Agency

Responding to that challenge, three of its research units: Embrapa Beef Cattle, Embrapa Information Technology and Embrapa Technology Transfer structured the Embrapa Information Agency project. The project was inspired in the ideas of the French anthropologist Pierre Lévy (Lévy, 1999), who proposed the use of Internet as a space for interaction and knowledge construction, instead of just merchandising, consumption and entertainment.

In the Information Agency, information is organized using two criteria. Firstly, every identified resource is cataloged in accordance to Dublin Core Metadata Standard (Dublin Core Metadata Initiative, 2003). Secondly, a set of terms belonging to the domain of discourse of farmers, rural extension technicians and researchers is identified.

Those terms were organized in the form of a tree and embedded in the system's navigational structure, which provides an option for hyperbolic tree visualization, illustrated in Figure 1. Since hyperbolic space provides an exponential amount of room, there is a natural balance between the exponential growth in the number of nodes and the space available as tree depth increases (Eick, 2001).

Moreover, the domain chosen for the first prototype system was the beef cattle supply chain. That choice was guided by the hypothesis that the supply chain is an adequate "ontology" for the exchange of knowledge among the intended users.

According to Holpsapple & Joshi (2002), an ontology is an explicit specification of an abstract view of a domain. In the Agency, the knowledge tree can be viewed as a *light* ontology, in the sense that it was built using the terms identified in the domain of discourse of the actors involved in the technology transfer activity and related them in a tree structure. However, for the reasons pointed in the next section, it has to be improved to satisfy the requirements for the Semantic Web Initiative.

Improving the model to encompass complete ontologies

Although extremely useful from the point of view of humans, the Information Agency is translucent as regards the computers. They can see a myriad of resources, all cataloged in Dublin Core, which are referenced by pages. Conversely, they cannot see the information inside the hyperbolic tree, e.g. why nodes B and C are subnodes of node A.

The hyperbolic tree is implemented in the form of an applet, which reads a compacted file containing all the nodes, its related information for displaying, and the linking among nodes, i.e. which node is the parent of another one. Search engines are not able to decode that compacted file, and so are not able to establish relations among the pieces of information.

In accordance to the Semantic Web Initiative, the higher the position in the stack displayed in Figure 2, the better is the quality of the information search engines and robots will be able to

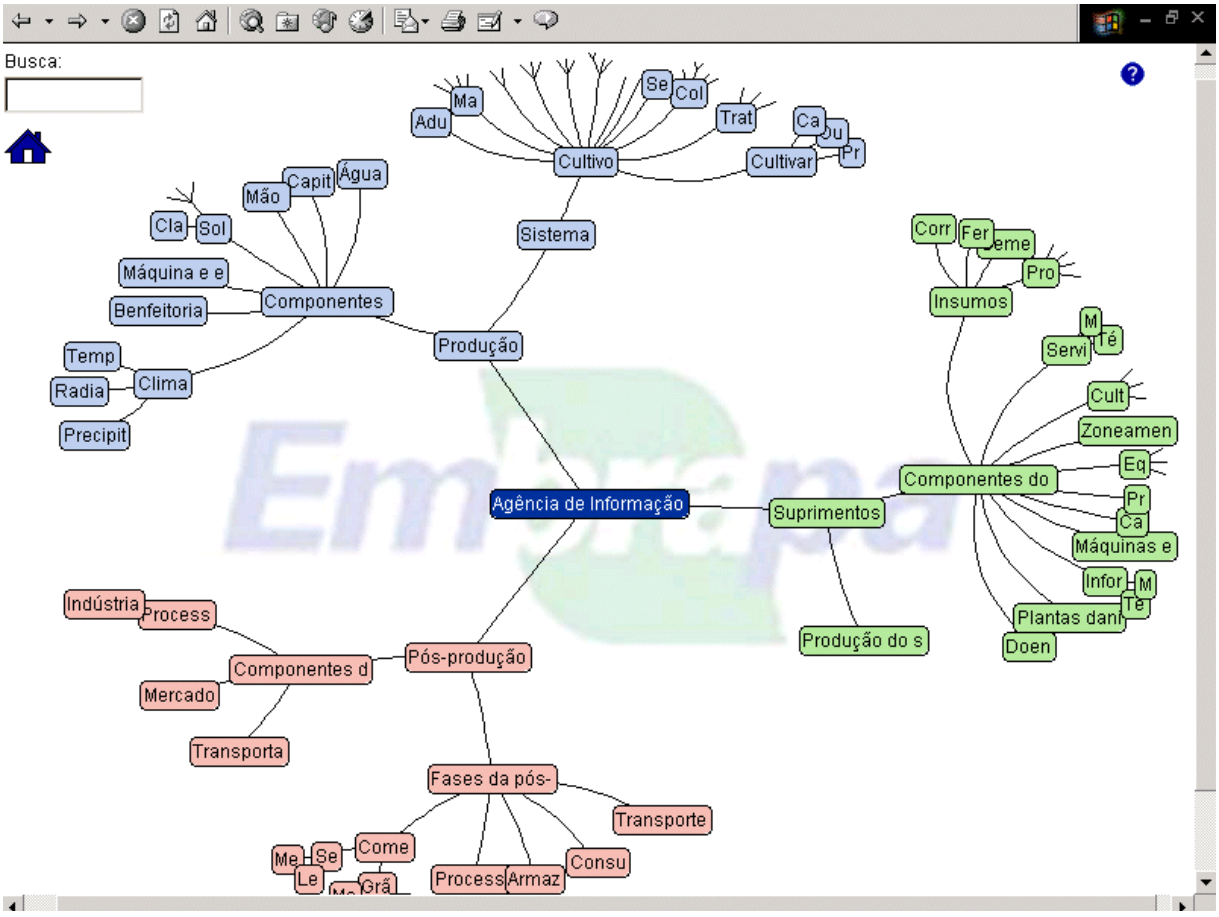

Figure 1: hyperbolic tree visualization of the beef cattle information agency

provide.

Examining the organizational structure of information for the beef cattle domain, it can be found that:

- Because each information resource is cataloged in accordance to Dublin Core Metadata Standard, regarding the cataloged resources, the specification went only up to the description of the resources (RDF), corresponding to the third layer in Figure 2;
- The representation of the beef cattle domain in a tree, by identifying the vocabulary and disposing it in a hierarchy, extended the specification towards the construction of an ontology (fourth layer in Figure 2). However, that extension is not complete for it lacks its RDFS description and the upper layers. The RDFS contains the necessary information for describing

properties and classes of RDF resources, with a semantics for hierarchies of such properties and classes;

- Nevertheless, to satisfy Semantic Web it is absolutely essential that a complete specification of ontologies for each domain currently in construction (beef cattle, dairy cattle, rice, soy, cotton, bean, eucalyptus, swine and sheep) be made. Ontologies for the agriculture domain are almost inexistent. FAO announced a project with such intent but there have been developed only prototypes of ontologies for fishery and bio-security domains (FAO, 2003).

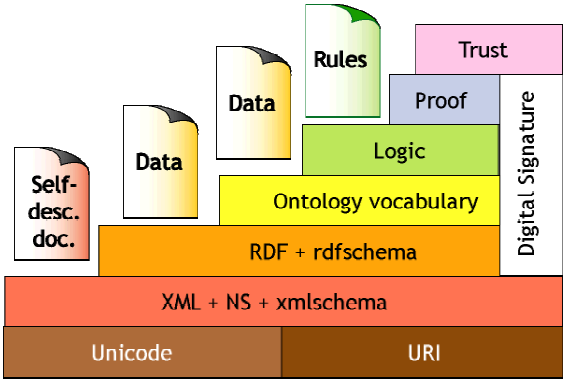

Figure 2: semantic web stack (Berners-Lee, 2000)

Embrapa´s initiative in splitting agricultural domain into subdomains and organizing information according to those subdomains is a good strategy towards an upper ontology for the whole agricultural domain. That has to be completed for the success of the project.

Another problem that many of the researchers working on a specific domain have pointed out is that there are some commonalities among certain domains, e.g. beef cattle and dairy cattle. Which suggests the necessity of studying the merge of ontologies at a certain level. For instance, the following facts belong to beef cattle, dairy cattle and sheep ontologies, respectively:

- cultivated_pasture is a kind of pasture that is eaten by beef cattle

  (is_a cultivated_pasture pasture)

  (eaten_by cultivated_pasture beef_cattle)

- cultivated_pasture is a kind of pasture that is eaten by dairy cattle

  (is_a cultivated_pasture pasture)

  (eaten_by cultivated_pasture dairy_cattle)

- cultivated_pasture is a kind of pasture that is eaten by sheep

  (is_a cultivated_pasture pasture)

  (eaten_by cultivated_pasture sheep)

These facts could be merged in the upper ontology resulting on simplified form:

(is_a cultivated_pasture pasture)

(eaten_by cultivated_pasture beef_cattle)

(eaten_by cultivated_pasture dairy_cattle)

(eaten_by cultivated_pasture sheep)

Therefore, the set of rules that defines an upper ontology encompassing the three domains should consider only the last set of rules, instead of the simple juxtaposition of the three other sets. Naturally, only humans can identify similar concepts with different names in different ontologies. In the previous cases, for instance, beef cattle ontology named *cultivated_pasture* what sheep ontology named just *cultivated*. Of course, that is clear for humans when they see the context but not for machines.

Suggestion of steps for improvement of the Agency model

The improvement of the agency model to satisfy the requirements identified above regarding ontologies and merge of ontologies could be performed according to the following lines:

1. study the current structure of the agencies and exercise their transformation into true ontologies, beginning with the beef and dairy cattle domains;
2. complement the ontologies developed with the logic layer, using logical languages such as OIL (Ontology Inference Language), DAML+OIL (unification of DARPA Agent Markup Language with OIL) (Connolly, 2001), or their upcoming successor OWL (Ontology Web Language) (Heflin, 2003);
3. verify the consistency of the logic layer using and inference engine such as FaCT (Fast Classification of Terminologies). FaCT is a Description Logic (DL) classifier, developed by the University of Manchester, that can be used together with Protégé2000 (Stanford Medical Informatics, 2003) or OilEd (http://oiled.man.ac.uk/). These are two of the intended ontology editors to be used in ontology construction;
4. study the ontologies of closely related domains and propose a merge method, as well as evaluate the possibility of existence of an upper merged ontology encompassing all domains. Merge of ontology is an open research subject in and that study is supposed to give some contribution for the case of closely related domains.

Conclusion

In this paper we presented the usage of the Information Agency as a space for exchange of knowledge among farmers, researchers and rural extension technicians. We suggested how the information contained in knowledge trees could be improved to become true ontologies in the sense of the Semantic Web Initiative, and presented common points identified in the knowledge trees already specified for different domains. These points indicated that there is a certain degree of merge among ontologies for different domains and that the proposition of an upper merged ontology for agricultural domain is feasible. A sequence of steps to reach full specified ontologies considering a possible merge among the existing ones was also suggested.

Bibliography

Berners-Lee, T. Semantic Web - XML2000 URI: http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide14-0.html. Accessed in February 2003.

CONNOLLY, D. et. al. DAML+OIL (March 2001) Reference Description. World Wide Web Consortium. URI: http://www.w3.org/TR/daml +oil –reference. Accessed in March 2003.

DUBLIN CORE METADATA INITIATIVE. Dublin Core metadata element set, version 1.1: reference description. URI: http://purl.org/dc/documents/rec-dces-19990702.htm . Accessed in February 2003.

EICK, S. G. Visualizing online activity. Communications of the ACM, v. 44, n. 8, p. 45-50, Aug. 2001.

FAO. FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS Agricultural Ontology Service Project (AOS). URI: http://www.fao.org/agris/aos/ .Accessed in March 2003.

HEFLIN, J. Web Ontology Language (OWL) Use Cases and Requirements. W3C Working Draft 3 February 2003. URI: http://www.w3.org/TR/ webont-req/ .Accessed in March 2003.

HOLSAPPLE, C. W.; JOSHI, K. D. A collaborative approach to ontology design. Communications of the ACM, v. 45, n. 2, p. 42-47, Feb. 2002.

HORROCKS, I. The FaCT System. URI: http://www.cs.man.ac.uk/~horrocks/FaCT/ Accessed in March 2003.

LÉVY, P. A inteligência coletiva: por uma antropologia do ciberespaço. 2. ed. São Paulo: Edições Loyola, 1999. 212 p.

PEIRCE, C.S. Semiótica. 3. ed. São Paulo: Editora Perspectiva, 2000. 337 p. (Estudos).

STANFORD MEDICAL INFORMATICS, The Protégé Project. http://protege.stanford.edu/ Accessed in March 2003.

RADA, R. Hypertext: from text to expertext. London: McGraw-Hill, 1991. 237 p.

SHOE. The SHOE FAQ. Parallel Understanding Systems Group. Department of Computer Science University of Maryland at College Park, USA. URI: http://www.cs.umd.edu/projects/plus/SHOE/ faq.html Accessed in February 2003

TENÓRIO, R. M. Cérebros e computadores: a complementaridade analógico-digital na informática e na educação. São Paulo: Escrituras Editora, 1998. 211 p.