PROCESS OF ONTOLOGY CONSTRUCTION FOR THE DEVELOPMENT OF AN INTELLIGENT SYSTEM FOR THE ORGANIZATION AND RETRIEVAL OF KNOWLEDGE IN BIODIVERSITY – SISBIO

Filipe Corrêa da Costa[1]; Hugo Cesar Hoeschl[1], Aires José Rover[1], Tânia Cristina D'Agostini Bueno [2]

[1] Research Institute on e-Gov, Juridical Intelligence and Systems – IJURIS, [2] WBSA Intelligent Systems S.A

{filipe, hudo, tania}@ijuris.org; aires.rover@gmail.com

http://www.ijuris.org

ABSTRACT: This work describes the ontology construction process for the development of an Intelligent System for the Organization and Retrieval of Knowledge in Biodiversity – SISBIO. The system aims at the production of strategic information for the biofuel chain Two main methodologies are used for the construction of the ontologies: knowledge engineering and ontology engineering. The first one consists of extracting and organizing the biofuel specialists´ knowledge, and ontology engineering is used to represent the knowledge through indicative expressions and its relations, developing a semantic network of relationships.

KEY WORDS: Ontologies, Knowledge Engineering, Ontology engineering, Bio-fuel

1. INTRODUCTION

The matter of energy has always been extremely important in the history of humanity. From craft production to the strong industrialization process, along with transportation and other uses, energy is an essential element for the development of the great powers of the world and is a limiting factor for the developing countries. For Sawin, "everything that we consume or use – our homes, their content, our cars and the highways on which we drive, our clothes and the food that we eat – require energy to be produced and packaged, delivered to the stores or homes, operated and then discarded" [1].

Market expansion caused by the strong industrialization process, the huge population growth and concentration, the military disputes, the forecasts of

future lack of natural resources and globalisation, lead to an excessive consume of energy, elevating its value considerably. These factors, increased by the predominant use of fossil fuels, cause severe economic, social and environmental problems, i.e. the greenhouse effect and climate changes.

In this context, the production of biofuel became an effective way to improve the diversification of the energetic matrix, contributing for the preservation of the environment and the economic and social development. For this, it is necessary to establish mechanisms of support to the production and commercialisation of biofuel. It will only be possible through the production of strategic information that will help in variables such as the optimisation of resources, integration of public and private institutions and rural producers. This information must be reliable, and studies must be made to guarantee the juridical, environmental, social, technical and commercial viability and a knowledge management process becomes necessary.

The use of correct information is one of the best ways to manage natural resources in a sustainable way, and is an important instrument in the decision-making processes. In a globalised world characterised by the excess of digital documents, information technology is rapidly evolving and is providing applications for many knowledge domains. In Brazil, although there are many scientific studies and some financing mechanisms, there is no structure able to provide reliable information and integrate the many agents that take part in the biofuel chain. Biofuels have differentiated productive chains and very specific characteristics. In the chain of biodiesel, for example, we must take into consideration factors related to agriculture, storage, characterization and quality control, co-production and, finally, commercialization and distribution themselves. Another example, is the natural gas chain, that goes through the stages of exploration, exploitation, production, processing, transportation, storage and distribution.

For the visualization and conception of SISBIO model- Intelligent System for the Organization and Knowledge Retrieval in Biodiversity, bio-business was defined in the area of biofuel as the development of fuel from natural renewable resources, bringing about economic, social and environmental benefits at a local, regional and global level. In order to be characterized as a bio-business, this product must have some type of relationship with some sector (biofuel), an element (raw material) and an asset (biodiesel or solar energy).

In this sense, the present work intends to demonstrate that it is possible, through the development of an intelligent system for the organization and retrieval of knowledge in biofuel, to ally the concepts of socio-economic development and environment preservation, through the foment of biofuel along its chain.

SISBIO has its linguistic development based on ontologies, providing a more efficient and precise retrieval of relevant information.

The term ontology, often used in philosophy, is used here as a knowledge representation structure that aims at the sharing of knowledge from a domain between people and systems. In this sense, ontologies are used to bring a common understanding of a certain domain through the relation between words or indicative expressions that represent a context. They are used in the construction of Knowledge Based Systems (KBS).

The construction of ontologies enables a better performance of the system, mainly in the precision and contextualization of the search results. Comparing to key-work search, a system that uses ontologies has a great advantage, mainly when longer texts are inserted in the search field. Using ontologies, we can affirm that the longer the text, the better the results are and in the case of SISBIO, it is possible to enter texts up to 15.000 words. This performance is achieved because when knowledge is represented based on ontologies, the system is able to search using the context of the entry text and not only key-words and logic expressions.

The area of application of this work is the construction of an ontologies network for the development of SISBIO, enabling the identification of potential business related to the biofuel chain.

The conceptual development of SISBIO is described in the second section. The methodologies and techniques applied in the construction of the ontologies are presented in section 3, and some conclusions are discussed in section 4.

2. INTELLIGENT SYSTEM FOR THE ORGANIZATION AND RETRIEVAL OF KNOWLEDGE IN BIODIVERSITY – SISBIO

SISBIO software –Intelligent System for the Organization and Retrieval of Knowledge in Biodiversity, provides institutions, rural producers and investors an immediate access to relevant information for the decision making process. The system is able to generate dynamic reports and extract hidden knowledge from a database. It uses a methodology that enables the organization of information stored in structured and non-structured bases, generating extremely relevant knowledge for governmental institutions, rural producers and investors. Apart from producing, integrating and processing a big amount of relevant information, it also creates a relationship network among the agents that constitute a bio-business.

SISBIO is a system developed to organize knowledge and produce strategic information for the monitoring and decision-making processes. It is based on Artificial Intelligence, Knowledge Engineering and Ontology techniques, which enables the integration of structured databases – i.e. a table with graphs – with results from the processing of non-structured data, such as reports, dissertations, and theses among others.

It was developed to take care of the informational and organizational needs of the knowledge of the biofuel chain agents, besides stimulating the relationship between them. Its structure can be coordinated by an agricultural cooperative, organization of enterprise entities, such as the National Confederation of Industries - CNI, or even, something that would be more strategic, to be coordinated by some ministry of the federal government.

With the development and implementation of SISBIO, it will be possible to foment public and private investments for the sustainable use of the natural resources. This model based on knowledge will assist in the creation of an only database, fed by diverse sources, turning information more agile and transparent, stimulating the development of bio-business.

To reach its goal, the system must be prepared to monitor some kinds of information, such as: the national and international financing sources, subjects related to the production and price of biofuel and carbon credits negotiated in stock markets around the world, among other pieces of relevant information.

SISBIO presents the modules of collection, storage, analysis and relationship, apart from the ontologies and informative notes publisher.

The collecting module is responsible for the monitoring of open digital sources. Intelligent agents of collection are created, who monitor specific digital targets. The automatic agents of collect the pre-defined information of each source information indicated by the specialists of the domain and stores them in the knowledge base of the system.

The storing module is responsible for the organization of the knowledge based on the semantic ontologies and their relations. Each new document enclosed in the system, either by the observers, or Informative Notes, is automatically indexed and stored in the same structure. This module prepares the information to be retrieved and analyzed in the analysis modules.

The analysis module is responsible for the textual and graphic search through matching between the case presented and the documents stored in the knowledge base. Besides documents of the base, the informative notes are retrieved, which are documents inserted manually generally containing strategic analyses and information.

The relationship module is the environment that allows the identification of chances of bio-business through the registry of technological innovations and research in development, agricultural producers besides investors and owners of agriculturable lands. The module is fed by the insertion of data of the agents that compose the biofuel chain, each one with their own peculiarities.

The ontology editor is the environment for the creation and registry of the ontologies and their relations through the identification of relevant expressions and those that represent the knowledge domain. It is part of the ontology engineering that will be described in item 3.2.

The informative note environment is prepared to insert documents in the knowledge base produced by analysts and specialists of the domain, generally containing relevant and strategic information for the generation of bio-business.

3. METHODOLOGIES AND TECHNOLOGIES APPLIED IN THE CONSTRUCTION OF SISBIO

SISBIO has its linguistic development based on ontologies. Two main methodologies are used for the construction of ontologies: knowledge engineering and ontology engineering. The first one consists of extracting and organizing the knowledge of the specialists in biofuel, and ontology engineering is used to represent the knowledge through indicative expressions and its relations, developing a semantic network of relationships.

3.1 Knowledge Engineering

Knowledge engineering is a methodology for the development of knowledge based systems. It is a fundamental step for the development of the model. This process is considered multidisciplinary by nature and includes some kinds of research that are difficult to classify in a delimited approach [2].

This process is responsible for the analysis of the knowledge domains that are used for the retrieval of information stored in the knowledge base used by SISBIO. Its also responsible for the requirement analysis, studies related to the biofuel chain, identification of information sources and definition of the interface.

The development stage of SISBIO initiates with the stage of Knowledge Engineering, responsible for the analysis of requirements, studies related to biofuel chain, identification of the informational sources, and definition of the fields of the interface. It is organized in the Knowledge Engineering suite. The suite consists of an independent computational structure for the extraction, organization and representation of the knowledge extracted in the phase of Knowledge Engineering, beyond the construction and edition of ontologies performed in the coming phase of Engineering of ontologies. The Knowledge Engineering Suite was developed to act together with Dynamically Contextualized Knowledge Representation- DCKR. Among the main tools, we can highlight the frequency extractor and the semantic extractor and the ontology editor.

An important step is the selection of information sources that will be used by the system. For SISBIO, the sources are institutions that provide information about biofuel, research organizations and public databases. The informational sources are proceeding from the organisms that possess

information about the biobusiness chain and research institutions, beyond public databases such as, the Brazilian Institute of Geography and Statistics - IBGE, Mines and Energy Ministry among others. It is through the analysis of documents found in these sources that the indices are defined, those that will be the base for the extraction of information. Those indices are defined to facilitate the process of retrieval of documents related to the consultation made by the user. This definition must be performed according to the relevance found in the content of the documents.

After that, the fields of the interface are defined that will be made available to the final user. This definition must undergo an analysis of criteria that involves the navigability, usability and ergonomics, based in norm ISO 9.241, and is described as "the capacity of a product to be used for specific users to achieve a specific goal with effectiveness, efficiency, and satisfaction in a specific context of use" [3].

As a multidisciplinary effort, knowledge engineering demands the participation of actors with different characteristics, as shown in figure 1:

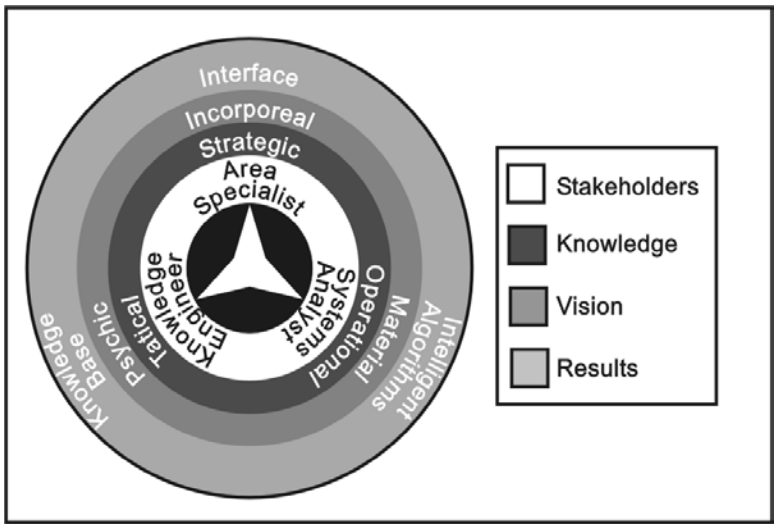

Figure 2: Knowledge Engeneering Actors

As it can be seen in the white area of the figure, there are three essential actors. The domain specialist is responsible for the organization and definition of the relevancy of the knowledge to be represented in the system knowledge base. Knowledge engineering is responsible for the extraction and representation of the knowledge of the specialist. Finally, the systems analyst is responsible for the definition of the tool adequate for the implementation of the requirements defined in this process.

3.2 Ontology Engineering

Ontology Engineering is a knowledge representation methodology that aims at the improvement of document retrieval performance. This representation is based on indicative expressions and the relations between them.

The use of ontologies enables the sharing of knowledge among people and the software, defining a common understanding about a defined knowledge domain. According to Studer et al. "an ontology is a formal, explicit specification of a shared conceptualization…Basically, the role of ontologies in the knowledge engineering process is to facilitate the construction of a domain model. Ontology provides a vocabulary of terms and relations with which a domain can be modeled" [4].

Systems that use ontology for document retrieval are less prone to risks of ambiguity of words or terms in the input text. These systems are able to "understand" the context of the search, returning better search results. Also, a special characteristic of SISBIO, the presentation of graphs based on ontologies, allows the user to interpret the results of the search in a temporal dimension.

The main steps in ontology engineering is the analysis of documents, domain visualization exercises, conceptual maps construction and inclusion of the ontologies in the system, using three tools called frequency extractor, semantic extractor and ontologies editor.

The frequency extractor is a tool used for the analysis of a set of documents, providing information about the frequency and distribution of words. This analysis also provides a statistical visualization, context analysis, and allows the verification of predominant subjects in the documents and the terminology used by the specialists of the domain.

The semantic extractor uses the knowledge base constructed with the ontologies editor and validates it comparing it to the content of the documents that were studied, allowing the reorganization of the textual knowledge map. It also allows the verification of the contexts related to the domain being considered. As a consequence of this analysis, it is possible to validate the vocabulary inserted in the ontologies editor and constructed through the frequency extractor.

The indicative expressions and its relations are added to the ontologies editor, creating a relationship tree considering the conceptual similarity of the terms and its relations and weights. It allows the storage of an associative structure of the knowledge of a domain. This kind of organization allows the dynamic attribution of weights to words or indicative expressions, related to the context of the search. It becomes a dynamic knowledge map.

It is after the first steps of the ontology engineering, the analysis of the documents, the definition of information sources and the interface, that the domains of the biofuel chain are defined, which will be addressed by the

system. These domains represent the main variables that indicate the need of production of relevant information inside a knowledge area.

The next step is the elaboration of the lists of indicative expressions and the relations among them. The indicative expressions are extracted by the specialists in biofuel and the knowledge engineers in the previous steps of the knowledge engineering. From these expressions, the vocabulary is expanded by means of technical and usual relations and by the validation done with the semantic extractor. Through the vocabulary constructed by the ontology engineering are found expressions analogous to the ones defined in the norm, as well as many other terms found in documents in a usual way. As a consequence, the relationship tree constructed from the ontologies editor "defines the linguistic, semantic and axiological similarity of conditions that allows the determination of local similarities between the values of an index" [BUENO, 2005].

In SISBIO, there were around 95 indicative expressions considered important, which generated around 415 new expressions after the expansion of the list of expressions with synonyms, related terms, "type of" and "part of". The following table exemplifies some of the main expressions.

List of indicative Expressions
Growth of agricultural potential
Sugar cane intercrop
growth of Biodiesel production
Program for the encouragement to alternative sources of electricity
Emissions exchange market

Table 1. List of indicative expressions

The classification of the ontologies is made through its relations and may be:

"Synonym" relation:

This relation exists between two terms or expressions that present the same meaning inside the same domain. Ex. Biofuel is a synonym of clean fuel.

"Related terms" relation:

The related terms relation occurs when the terms or expressions represent a strong relation, but do not have exactly the same meaning and do not have any other relation. Ex: "Exploitation taxes" is related to royalties.

Relation "Type of":

The type of relation is based on the classification between class and subclass. Ex. Biodiesel is a type of biofuel.

Relation "Part of":

This type of relation refers to the matter of hierarchies or organizational levels. Ex. Alternative energy is part of PROINFA.

Figure 2 represents the ontology editor interface, based on the classification above.

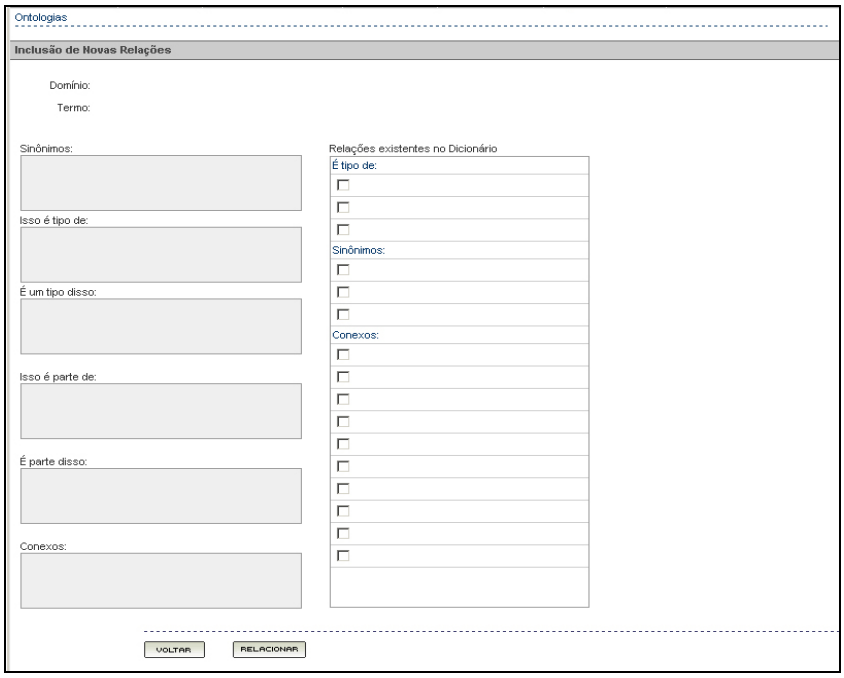

Figura 2 – Ontologies Editor
Fonte: Bueno, 2005

The ontology construction methodology consists of the following steps: 1. To produce an inventory of the whole domain, i.e. to catalogue all digital information sources that will be used as the database of the system; 2. to apply the frequency extractor in this database; 3. To make a comparison between the results of the extractors and the needs of the specialists; 4. To construct (with the aid of the specialist) a controlled vocabulary that represents the domain; 5. To apply the semantic extractor to the database, using the vocabulary; 6. To valuate the result based on the frequency of the indicative expressions found and define a list of words; 7. To construct ontologies to be used in the system based on this controlled vocabulary; 8. To define synonyms, homonyms and hyperonyms based on doctrines and the legislation [BUENO 2005].

4. CONCLUSION

The use of ontologies in informational systems based on knowledge is demonstrating a strong differential in terms of efficiency of the search results. The organization of knowledge through the construction of a expressions network allows the system to have documents as results organized by context and similarity degree in relation to the input document. It is also promoting a conceptual evolution in search engines, where contextual search and knowledge representation become essential elements for the production of strategic information. While in key-word and logical connectors based systems exists a strong ambiguity in the search results, in ontology based systems the context allows a better understanding of the situation presented by the input text, ranking the documents by relevancy and similarity.

Knowledge organization in domains is another differential, in the same way that specific information of each domain adopts particular contexts, defining the universe of scanlines of the system. Another relevant fact is that informational sources themselves that will be feeding the system are defined in a way to cover the whole domain.

The automatic construction of ontolgies is a challenge to be implemented in the system. And the improvement of the process for the construction and expansion of the ontologies is seen as a future work.

REFERENCES

1. SAWIN, Janet L.; Estado do Mundo, 2004: estado do consumo e o consumo sustentável / Worldwatch Institute ; apresentação Enrique Iglesias ; tradução Henry Mallett e Célia Mallett. - Salvador, BA : Uma Ed., 2004

2. BUENO T. C. D et al. Knowledge Engineering Suite: a Tool to Create Ontologies for Automatic Knowledge Representation in Knowledge-based Systems. in: The 2nd International Workshop on Natural Language Understanding And Cognitive Science (NLUCS-2005) in ICEIS - 7th international conference. 7TH International Conference on Enterprise Information Systems, 2005, Miami. Proceedings of Seventh International Conference On Enterprise Information Systems. 2005.

3. DIAS, Claudia Augusto. Métodos de avaliação de usabilidade no contexto de portais corporativos: um estudo de caso do Sanado Federal. 2001. 225 p. Tese (Doutorado) – Faculdade de Estudos Sociais Aplicados, Universidade de Brasília, Brasília/DF.

4. STUDER, R. et al., Situation and Perspective of Knowledge Engineering. In: J. Cuena, et al. (eds.), Knowledge Engineering and Agent Technology. IOS Press, Amsterdam, 2000.