A STRATEGY FOR BIODIVERSITY KNOWLEDGE ACQUISITION BASED ON DOMAIN ONTOLOGY

A. C. F. Albuquerque[1,2], J. L. Campos dos Santos[2], J. F. de Magalhães Netto[1]
[1] *Universidade Federal do Amazonas - UFAM/DCC/PPGI – Manaus, AM – Brazil*
[2] *Instituto Nacional de Pesquisas da Amazônia – INPA/PPBio/NBGI – Manaus, AM – Brazil*
andreaa@inpa.gov.br, lcampos@inpa.gov.br, jnetto@dcc.ufam.edu.br

Abstract

Convention on Biological Diversity (CBD) recognizes that biodiversity loss must be reduced to promote poverty alleviation and direct benefit of all live on Earth. To achieve that, we must consider robust strategies and action plans based on knowledge and state of art technology. Parallel to that, research is underway in universities and scientific organization aiming to develop semantic web as an additional resource associated to formal ontology and the avoidance of knowledge acquisition problems such as expertise dependence, tacit knowledge, experts' availability and ideal time importance. Ontology can structure knowledge acquisition process for the purpose of comprehensive, portable machine understanding and knowledge extraction on the semantic web environment. These technologies applied to biodiversity domain can be a valuable resource for CBD. The paper presents a strategy for biodiversity knowledge acquisition based on a negotiation protocol which uses domain ontology to extract knowledge from data sources in the semantic web domain.

1. Introduction

Biological diversity, or biodiversity, is the term given to the variety of life on Earth. It is the combination of life forms and their interactions with one another, and with the physical environment that has made Earth habitable for humans. Ecosystems provide the basic necessities of life, offer protection from natural disasters and disease, and are the foundation for human culture [1].

To concern over the loss of biodiversity and the recognition of its important role in supporting human life motivated the creation, in 1992, of the CBD, a legally binding global treaty [2]. The Convention encompasses three equally important and complementary objectives: the conservation of biodiversity, the sustainable use of its components, and the fair and equitable sharing of benefits arising out of the utilization of genetic resources.

A particular resource to promote these effort is knowledge, which is intrinsic to people - mix of information, analysis, integration, experience, culture, etc. Much of knowledge is tacit - we cannot communicate our knowledge completely in words or symbols (e.g. the expertise of a craftsman). When we express our knowledge, it becomes explicit knowledge - what we can express/ articulate to others. Due to its characteristics to capture and manage knowledge is difficult but essential for all domains.

In order to succeed, CDB, should consider semantic web and ontology applied to biodiversity domain as powerful tools, aiming knowledge acquisition and integration.

This paper is organized as followed: Section 2 presents an overview of related works. Knowledge acquisition in perspective is presented in section 3. The strategy for knowledge acquisition is described in section 4. Finally, in section 5 the conclusions are presented.

2. Related Work

Unlike the existing web, where data is primarily intended for human consumption, the semantic web will provide data that is also machine processable. This will enable a wide range of intelligent services such as information brokers, search agents, information filters, etc., a process that Berners-Lee describes as "Bringing the web to its full potential" (Berners-Lee, Hendler, and Lassila, 2001). The idea is to add semantics to web

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.87

content in order to make it easier to find and use for both humans and machines. Adding formal semantics to the Web will help from resource discovery to the automation of wide range of processes.

The new generation of the Web highlights not only the benefits that come along, but also some drawbacks that appear in this new scenario. Finding data in large and heterogeneous digital collections (as in specific domains), especially on the Web, is increasingly difficult. One of the most frequently identified problems is how to search for and retrieve needed information from the large number of information sources available on the Internet. The information provided by the sources is no longer just simple text, but now includes multimedia, forms, structured data, and executable code - it has become much more complex. As a result, old methods for manipulating these sources are no longer appropriate or even efficient. To keep pace with the complexity and growth of the Internet, it is necessary not only suitable storage and retrieval and knowledge management mechanisms, but also efficient search tools that can harvest the needed information from these sources.

Ontology is a specification of a conceptualization, that is, a description of concepts and relations that can exist for an agent or an agent community [4]. Ontology is (meta) data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine processable semantics. By defining shared and common domain theories, ontology helps both people and machines to communicate concisely, supporting the exchange of semantics and not only syntax. They will therefore have a crucial role in enabling content-based access, interoperability and communication across the web. Hence, the light and fast construction of domain-specific ontology is crucial for the success and the proliferation of the semantic web.

One of the main benefits of using ontologies, whatever is the scenario, is to reuse domain specifications in the requirement specification phase. Reusable ontologies are becoming increasingly important for tasks such as information integration, knowledge-level interoperation and knowledge-base development.

Using ontology, the effort for knowledge acquisition can be divided into two phases: (1) Explicit specifications of the basic conceptualization of the domain are created in the form of ontology, having as focus the common knowledge on the domain, common to an ample set of applications; (2) The specific knowledge of an application is captured and codified in a Knowledge-Based System (KBS). The later is strongly guided by ontology, since these provide to the knowledge engineer a vocabulary to express the domain, through the terms of the ontology, and a nucleus of knowledge, supplied for its axioms. We consider that an advantage of using ontology in the development of KBSs: when we divide the knowledge acquisition into two phases, we have a practical and objective way to reuse and to share knowledge bases and to offer valuable lines of direction for the orientation of the acquisition of the specific knowledge of an application.

In a distributed and heterogeneous environment such as the Internet, ontology-based manipulation of these diverse sources is a useful solution to guide the knowledge acquisition.

Knowledge discovery from documents, such as Web pages, is a useful but still complex task. Ontologies can achieve a high degree of accuracy in knowledge acquisition while maintaining resiliency faced in document changes.

Since the main obstacle to knowledge management comes from the lack of a basic conceptualization of the domain that will be used, ontology emerges as the key point in the search of knowledge acquisition [5]. Domain ontology must be developed and become available, in the form of a modular knowledge base, to be used to guide the acquisition of the desired specific knowledge, also allowing the reuse and the sharing of knowledge.

3. Knowledge Acquisition in Perspective

Knowledge Acquisition is the transformation of knowledge from the forms in which it is available in the world into forms that can be used in a knowledge base. A knowledge base is a centralized repository for information: a public library, a database of related information etc., or even, in relation to information technology, is a machine-readable resource for the dissemination of information, generally online or with the capacity to be put online.

Knowledge acquisition includes the elicitation, collection, analysis, modeling and validation of knowledge for knowledge engineering and knowledge management projects.

Some of the most important issues in knowledge acquisition are as follows: Most knowledge is in the heads of experts; Experts have vast amounts of knowledge; Experts have a lot of tacit knowledge (They don't know all that they know and use; Tacit knowledge is hard to describe.); Experts are very busy and valuable people; Each expert doesn't know everything; Knowledge has a "shelf life".

It’s usually not possible to transfer a domain’s expert knowledge directly to a KBS because the respective representations of knowledge are too dissimilar. Domain experts often explain their knowledge thought the use of anecdotes and examples, but for a KBS more general principles are required. Experts frequently provide incomplete and even incorrect knowledge, or may not be able to articulate their knowledge at all. Experts may not have the required attitude to communicating their knowledge, or insufficient time or resources to do so properly. Also multiple experts may have significantly differing opinions on what is or is not the right or wrong way to do things.

Because of these issues, techniques are required which: take experts off the job for short time periods; allow non-experts to understand the knowledge; focus on the essential knowledge; capture tacit knowledge; allow knowledge to be collated from different experts; allow knowledge to be validated and maintained.

3.1. Knowledge Acquisition Techiniques

Many techniques have been developed to help elicit knowledge from an expert. These are referred to as knowledge elicitation or knowledge acquisition (KA) techniques. The term "KA techniques" is commonly used.

The following list gives a brief introduction to the types of techniques used for acquiring, analyzing and modeling knowledge: Protocol-generation techniques include various types of interviews (unstructured, semi-structured and structured), reporting techniques (such as self-report and shadowing) and observational techniques; Protocol analysis techniques are used with transcripts of interviews or other text-based information to identify various types of knowledge, such as goals, decisions, relationships and attributes. This acts as a bridge between the use of protocol-based techniques and knowledge modeling techniques; Hierarchy-generation techniques, such as laddering, are used to build taxonomies or other hierarchical structures such as goal trees and decision networks; Matrix-based techniques involve the construction of grids indicating such things as problems encountered against possible solutions. Important types include the use of frames for representing the properties of concepts and the repertory grid technique used to elicit, rate, analyze and categorize the properties of concepts; Sorting techniques are used for capturing the way people compare and order concepts, and can lead to the revelation of knowledge about classes, properties and priorities; Limited-information and constrained-processing tasks are techniques that either limit the time and/or information available to the expert when performing tasks, and; Diagram-based techniques include the generation and use of concept maps, state transition networks, event diagrams and process maps. The use of these is particularly important in capturing the "what, how, when, who and why" of tasks and events.

The various techniques described above shows the types of knowledge they are mainly aimed at eliciting. Visualizing them in a graphic, the vertical axis would represents the dimension from object knowledge to process knowledge, and the horizontal axis, the dimension from explicit knowledge to tacit knowledge.

Ontology can be used as a KA technique able to specify knowledge making use of the available KA techniques. The main use of ontology is to share and communicate knowledge, both between people and between computer systems. A number of generic ontologies have been constructed, each having application across a number of domains which enables the re-use of knowledge.

In this research context, we aim to make use of biodiversity domain ontology in the semantic web in order to provide mechanism for knowledge acquisition. Figure 1 illustrates this scenario.

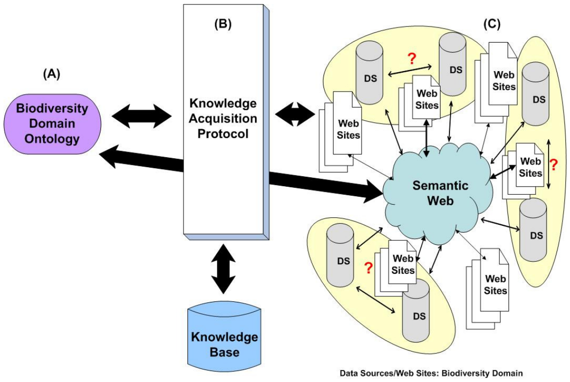

(A)
Biodiversity Domain Ontology
(B)
Knowledge Acquisition Protocol
Knowledge Base
(C)
DS
Web Sites
?
Semantic Web
Data Sources/Web Sites: Biodiversity Domain

Figure 1. Architecture to extract knowledge from the web.

4 The Strategy for Knowledge Acquisition

4.1. Biological Domain Ontology

Biodiversity information is the chosen application domain. Once it is a large and complex domain, the objective, among others, is to verify if ontology can satisfactorily be used as an approach for biological knowledge acquisition.

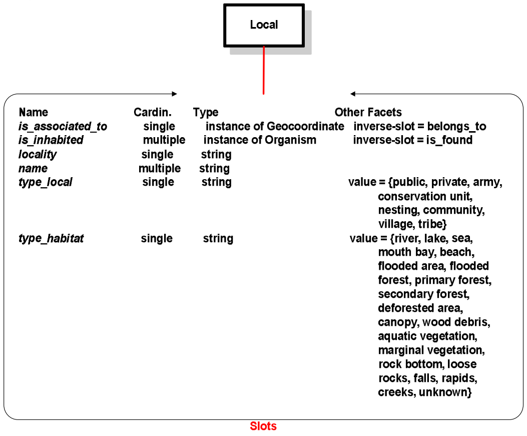
At the very beginning of this research there was not available any specific work to guide biodiversity data modeling. The solution was search for every/any type of documents that belonged to the domain and define a generic type of document. Even with a very hard effort it probably would never cover the whole domain ontology modeling, updates need to be done constantly.

Campos dos Santos [6], concluded a research in biodiversity informatics and presented a schema representation of INPA's collections, named CLOSi (Clustered Object Schema for INPA's Biodiversity Data Collections) that can be the basis for an integrated view of its biological collections data. CLOSi is the result of a study carried out in selected scientific institutions in the Amazon and in other parts of the world. CLOSi and the material previously acquired is the base for the ontology developing.

To assist users in developing and maintaining ontologies a number of tools have been developed and a few comparative studies of ontology tools have been performed. Based on a survey [7], Protégé 2000/OWL [8] was the chosen tool to build the domain ontology.

Protégé-2000 is a tool for creating, editing and browsing ontologies developed by Stanford Medical Informatics. The design and development of Protégé-2000 has been driven by two goals: to be compatible with other systems for knowledge representation and to be an easy to use and configurable tool for knowledge extraction.

For the purposes of this research, as illustrate in Figure. 2 an ontology is a formal explicit description of concepts in a domain of discourse (classes, represented by the square in black line), properties of each concept describing various features and attributes of the concept (slots, called roles or properties, dotted lines balloons), and restrictions on slots (facets, role restrictions, specified for each slot). Ontology together with a set of individual instances of classes constitutes a knowledge base.

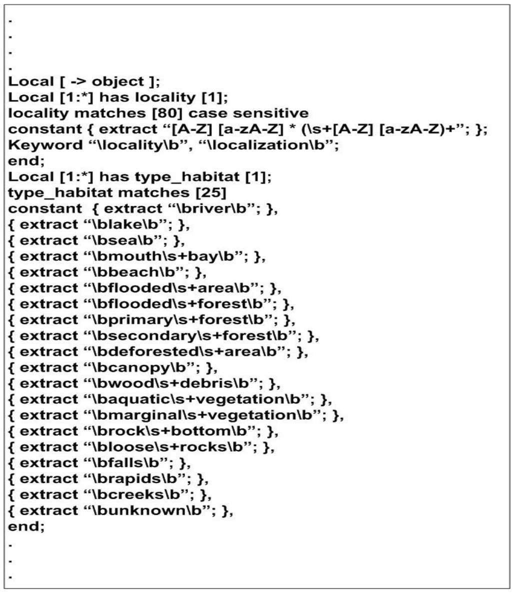


Name	Cardin.	Type	Other Facets
is_associated_to	single	instance of Geocoordinate	inverse-slot = belongs_to
is_inhabited	multiple	instance of Organism	inverse-slot = is_found
locality	single	string	
name	multiple	string	
type_local	single	string	value = {public, private, army, conservation unit, nesting, community, village, tribe}
type_habitat	single	string	value = {river, lake, sea, mouth bay, beach, flooded area, flooded forest, primary forest, secondary forest, deforested area, canopy, wood debris, aquatic vegetation, marginal vegetation, rock bottom, loose rocks, falls, rapids, creeks, unknown}

Figure 2. Partial ontology for biodiversity domain data sources (object Local).

Classes are the focus of most ontology and describe concepts in the domain. For example, a class of Geocoordinates represents all geographic coordinates. Specific coordinates are instances of this class. A class can have subclasses that represent concepts that are more specific than the superclass. For example, we can divide the class Organism into Microorganism, Flora, and Faun. Slots describe properties of classes and instances. At the class level, we can say that instances of the class Organism will have slots describing their genus, nick name and specie.

In practical terms, developing an ontology (using Protégé) includes: defining classes in the ontology; arranging the classes in a taxonomic (subclass–superclass) hierarchy; defining slots and describing allowed values for these slots and filling in the values for slots for instances.



```
.
.
.
.
Local [ -> object ];
Local [1:*] has locality [1];
locality matches [80] case sensitive
constant { extract “[A-Z] [a-zA-Z] * (\s+[A-Z] [a-zA-Z)+”; };
Keyword “\locality\b”, “\localization\b”;
end;
Local [1:*] has type_habitat [1];
type_habitat matches [25]
constant  { extract “\briver\b”; },
{ extract “\blake\b”; },
{ extract “\bsea\b”; },
{ extract “\bmouth\s+bay\b”; },
{ extract “\bbeach\b”; },
{ extract “\bflooded\s+area\b”; },
{ extract “\bflooded\s+forest\b”; },
{ extract “\bprimary\s+forest\b”; },
{ extract “\bsecondary\s+forest\b”; },
{ extract “\bdeforested\s+area\b”; },
{ extract “\bcanopy\b”; },
{ extract “\bwood\s+debris\b”; },
{ extract “\baquatic\s+vegetation\b”; },
{ extract “\bmarginal\s+vegetation\b”; },
{ extract “\brock\s+bottom\b”; },
{ extract “\bloose\s+rocks\b”; },
{ extract “\bfalls\b”; },
{ extract “\brapids\b”; },
{ extract “\bcreeks\b”; },
{ extract “\bunknown\b”; },
end;
.
.
.
```

Figure 3. Partial code of object Local of the biodiversity domain ontology.

Once, the web semantic can be seen as a huge database, in other words, a universal repository, all kind of data sources can be found. A domain ontology appears to help users (man and machine) find and get the desired data. It integrates all data sources in a unique base (semantically) based on domain ontology.

Figure 3 represents the code (using regular expressions from Perl) based in the domain ontology developed, used to identify and acquire knowledge from the different data sources. To do so, as strategy, we make use of a Knowledge Acquisition Protocol.

4.2. Knowledge Acquisition Protocol

Systems that interoperate about biological data must be able to integrate the different data sources found in a web semantic context and providing knowledge acquisition through the generation of a useful knowledge base. A protocol to manage knowledge acquisition can address the problem of integrating data sources found in the semantic web through the creation of knowledge bases, providing knowledge acquisition and using domain ontology for that, as illustrated in Figure 4.

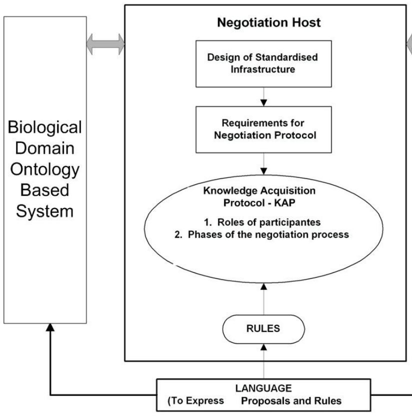

Negotiation Host
Design of Standardised Infrastructure
Requirements for Negotiation Protocol
Knowledge Acquisition Protocol - KAP
1. Roles of participantes
2. Phases of the negotiation process
RULES
LANGUAGE (To Express Proposals and Rules
Biological Domain Ontology Based System
Biological Data Sources (WEB)

Figure 4. Framework for knowledge acquisition.

The problem that we want to address is the following: An information service (ontology based system) R (requester) intends to request information from another service (information sources) P (provider) related to biological knowledge. To this purpose, it needs to identify one or more integration meaning points, and ensure some level of guarantee of semantic agreement, that is, the two systems try to ensure that they agree on the knowledge under discussion.

Two simplifying assumptions: the communication between R and P concerns about an explicit knowledge K; system R resolves issues of ontology rules and descriptions;

The problem at hand is a semantic agreement problem: how does system R indicate to system P which knowledge unit it is filing its request about? It does through the domain ontology developed.

4.2.1 A Communication Protocol

Several issues need to be considered for implementation. One of them is the need for a negotiation set-up that enables independent knowledge representation to interact using a form of negotiation. This set-up would cover aspects including defining a protocol for negotiation (definition of requester and provider, roles and phases of negotiation); defining language(s) for negotiation rules and to express negotiation proposals for the requester and for the provider.

Negotiation protocol determines the flow of messages between requester/provider, controlling the agents messages, when and actions according to the rules by which the negotiation process must follow.

The design of a standardized set-up is needed to allow independent interactions over time to reach agreement. A framework for negotiation between two agents should have the following characteristics: agents must be free to have their own infrastructure, providing pre-requisite functions to support for the automation of the negotiation process; the infrastructure should enforce the standardization of basic interaction rules; the protocol must guarantee that no party has access to extra information or to be able to forge false knowledge.

Figure 5 illustrates the communication protocol process. This process consists of the following steps:

1. R → P Full knowledge request, depending on ontology specification, this could for instance be a knowledge extraction in the semantic web; 2. R ← P Acknowledgement (or negative acknowledgement), indicating that P is (not) willing to accept the request, and in affirmative case is prompting for additional information, regarding knowledge (a knowledge unit or more) mentioned in full knowledge request; 3. R → P Initial request, providing relevant knowledge; 4. R ← P Initial reply, which will include a proposed definition of the content, from perspective of P; 5. R → P Optional additional request, indicating the content that R finds missing in P's initial reply. This constitutes a request for more data; 6. R ← P Additional reply, providing added answers to the initial reply given earlier; 7. R → P Full knowledge request, upon selection by R of elements earlier proposed by P; 8. R ← P Full knowledge reply by P taking into account the high level query earlier submitted by R.

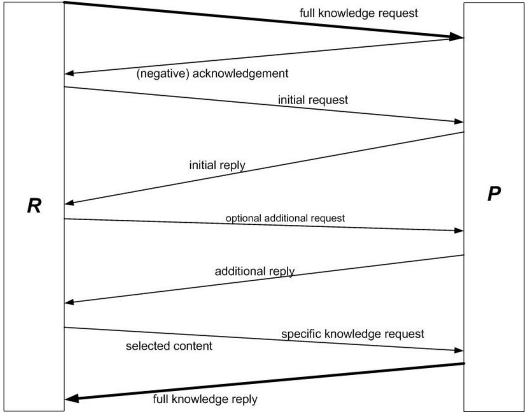


Figure 5. The communication protocol.

4.2.2 Rules and Negotiation Proposals

The format to express negotiation proposals has to be standardized. A language to attend the basic requirements must take into account the following: support for ontology and namespace; high degree of expressiveness; ability to express fully bound specifications; ability to express constraints over all possible values; loose support of types and inheritance; support for complex queries; and support for complex matching.

Table 1, presents a summarized language feature scope covering XML, RDF, and OWL. A brief discussion about these features is presented in DAML. – ORG.

The language OWL appear to be better suited to deliver the domain ontology proposals, rules and the communication protocol.

Table 1. Language feature comparison. (source: DAML – ORG)

Features	XML DTD	XML Schema	RDF (s)	OWL
bounded lists	x	x		x
cardinality				x
class expression				x
data types		x		x
defined classes				x
enumerations	x	x		x
equivalence				x
extensibility				x
formal semantics				x
inheritance			x	x
inference				x
local restrictions				x
qualified constraints				
reification			x	x

5. Conclusions

Despite the importance of ontology applied to a specific domain description, to build them is a challenging. The task comprises of the specification of concepts and relations that may exist in the domain, besides their definitions, properties and constraints. Therefore, all kinds of support for ontology development are welcome.

A major obstacle is the lack of guidelines on what knowledge parameters such ontologies should contain and what design principles they should follow. In the biodiversity informatics scope, for example, different communities use different approaches to build specific ontologies for semantically describe the same domain. Further, to build a generic and rich domain ontology one would ideally need to inspect a large number of web services and specific documents of that application area. Therefore formal support for ontology will help curators to get a quick insight in these large and dynamic data sets. Additionally, books, papers, manuals, web pages and other literature are very important sources for modeling ontology, but they are not enough. It is necessary to reach consensus among expert's specifications regarding specific domains.

Knowledge acquisition is a necessary technology to deal with the huge and growing collection of growing knowledge placed in the web semantic. Ontology-based knowledge acquisition is a robust approach, but the design of ontology is a technical task requiring the services of human expertise.

The use of a graphical language for expressing ontology proved to be essential for capture ontology concepts. It is very difficult to communicate with domain experts without graphical language. Languages for ontology textual representation must be faced as a delicate issue. Most of the languages used, are still not well adequate for this purpose.

Although the knowledge-acquisition-ontology approach is successful for many applications, one critical difficulty remains: a user must create the extraction ontology manually, which is both time-consuming and error-prone. Furthermore, knowledge-acquisition-ontology requires a high degree of knowledge in both database theory and Perl regular expressions.

The proposed protocol enables biological knowledge acquisition allowing the integration of data sources found in the semantic web through the creation of knowledge bases, providing knowledge acquisition and using domain ontology for that.

6. References

[1] Grime, J. P.: *Biodiversity and Ecosystem Function: The Debate Deepens*. Science Vol. 277 n^{o}.533029, pp. 1260 – 1261. (2007).

[2] Secretariat of The Convention on Biological Diversity: *Global Biodiversity Outlook 2*. 81 + vii pages. Montreal. ISBN-92-9225-040-X. (2006).

[3] Berners-Lee, T.; Hendler, J.; Lassila, O.: *The Semantic Web*. Scientific American 284 (5): 34-43. (2001).

[4] Guarino, N.: *Understanding, Building and Using Ontologies: A Commentary to Using Explicit Ontologies in KBS Development*. International Journal of Human and Computer Studies, v.46, n.2/3, p. 293-310. (1997).

[5] Falbo, R.: *Integração de Conhecimento em um Ambiente de Desenvolvimento de Software*. Tese de Doutorado. Programa de Engenharia de Sistemas e Computação da COPPE/UFRJ, Dezembro, (1998).

[6] Campos Dos Santos, J. L.: *A Biodiversity Information System in an Open Data/Metadatabase Architecture*. Ph. D. Thesis. International Institute For Geo-Information Science and Earth Observation. Enschede, The Netherlands. ISBN 90-6164-214-0. (2003)

[7] Lambrix, P., Habbouche, M., Pérez, M.: *Evaluation of Ontology Development Tools for Bioinformatics*. Bioinformatics, Vol. 19, no. 12 2003, p. 1564 - 1571. (2003).

[8] Noy, N., Sintek, M., Decker, S., Crubézy, M. and Musen, M.: *Creating Semantic Web Contents with Protege-2000*. In IEEE Intelligent Systems, PP. 60 – 71. (2001).