OntoBio: Designing New Features to Improve Modeling and Implementation

Andréa Corrêa Flôres Albuquerque[1], José Laurindo Campos dos Santos[2], Alberto Nogueira de Castro Júnior[1]

[1] Instituto de Ciência da Computação (IComp), Universidade Federal do Amazonas (UFAM), Manaus-AM, Brasil
[2] Laboratório de Interoperabilidade Semântica (LIS), Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus-AM, Brasil

andreaalb.1993@gmail.com, laurindo.campos@inpa.gov.br, alberto@icomp.ufam.edu.br

Abstract

OntoBio is a formal ontology developed in the scenario of biological collection and field data collection of biotic entities. Considering the complex and dynamic nature of biodiversity data and information, modeling and implementations decisions likely to be error prone, can happen. This paper presents OntoBio's limitations regarding conceptualization and implementational aspects and new features aiming to indicate accurate recommendations for OntoBio's evolution, by emphasizing several aspects that must be considered when designing a new version of the ontology.

1. Introduction

The current research on data integration has focused on semantics, which aims to mitigate the conflicts between heterogeneous data sources instead of designing the structure of an architecture for integration.

One strategy that has been adopted to deal with such problems is the use of integrative elements - such as ontologies - to manage and eliminate semantic conflicts. In the scope of biodiversity data and information, ontologies can be a valuable resource for strategic planning and contribution toward conservation [Albuquerque et al., 2015]. There is a remarkable growing demand for this data in several applications, such as environmental impact assessment, definition of environmental preservation areas, protection of endangered species, land reclamation, bio-prospecting, setting public policy, environmental legislation, among others.

Due to the wide-ranging characteristics of data and the diverse profiles of experts, there is still much work to be done in the specification of ontology for this domain. This is one of the reasons that the integration of biodiversity data and ecological studies is not considered trivial. Solutions for interoperability are needed for research in this field.

Regarding these facts, OntoBio, a formal ontology applied to biodiversity data, provided important results with already validated technology for the adoption of formal ontologies to knowledge acquisition and integration in biodiversity field. OntoBio was developed in a research initiative involving the IComp/UFAM and INPA's Biological Collection Program. It was modeled conceptually through OntoUML language [Guizzardi 2005] and developed through the SABIO method [Falbo et al. 1998]. OntoBio is divided into five sub-ontologies, connected by relationships between concepts and axioms. They are: collection[1]; material entity, that is composed by biotic entity and abiotic entity; spatial location; ecosystem; and environment [Albuquerque et al., 2015].

Considering the complex and dynamic nature of biodiversity domain, it is expected the occurrence of extensions/evolution of ontology, according to the views of experts. Elicited requirements with researchers from INPA guided the identification of new entities, categorizations, relationships and some new sub-ontologies. During the development of OntoBio, much of an expert's knowledge (which was not presented in the structured databases that support the ontology) was not represented, and thus lost. Empirical evidence indicated that this knowledge could become essential to incorporate semantic expressiveness in ontologies. A conceptual framework was proposed to aggregate scientific tacit knowledge into ontologies [Albuquerque et al., 2016]. The new version of OntoBio incorporates more semantics to the model and the availability of a version with features that allow its use in more complex applications (taxonomic classification).

2. OntoBio's New Features

Alloy language has been used as a way to evaluate graphic models, aiding the professionals that build them [Jackson, 2002]. OntoUML is a well-founded language to build ontologies. The existence of algorithms that translate models developed in this language to Alloy specifications helped the validation of OntoBio.

Due to the complexity of OntoBio and the size restriction (number of classes modeled) imposed by Alloy, it was validated in a segmented way: the strongly connected sub-ontologies were validated first, followed by the intersections of these sub-models. The validation identified recurrent modeling decisions that are error prone and they were presented in [Sales 2012].

In addition to the validation and suggestions of improvements found in [Sales 2012], some conceptual modeling aspects were considered:

- *Collection*. This sub-ontology can be segmented in two sub-ontologies: *Acquisition* and *Research Institution*;
- *Acquisition*. A Collection is defined as the acquisition of an organism, animal, vegetal, fungal or microbial. In *Acquisition*, the *Collection* entity is called

[1] Collection here means the act of collecting material entities in an environment.

Expedition, which is one of the ways of acquiring specimens. Other forms that must be considered are: *Purchase, Donation, Legacy and Exchange*. The collections performed by an *Expedition* must follow specific collection protocols.

- *Research Institution*. Currently the *Research Institution* has a broader representation, where additional features can be incorporated. For official Brazilian institutions, a biological collection comprises of properly treated biological material, maintained and documented in accordance with norms and standards to ensure the safety, accessibility, quality, longevity, integrity and interoperability of data collection, belonging to the scientific institution in order to support scientific or technological research and *ex situ* conservation.
- *Ecosystem*. It can be absorbed by the *Environment* sub-ontology, as well as phytophysiognomy, vegetation and climatic region modules of the *Spatial Location* sub-ontology.
- *Environment*. New specializations of micro and macro environment must be added to the *Environment* sub-ontology.
- *Material Entity*. In the first version of OntoBio, this sub-ontology captured the taxonomic ranks of *family*, *genus* and *species*. The complete taxonomic classification of the organism is required, which results in the creation of the sub-ontology *Taxonomic Classification* for this purpose. Food habits and maturity stage can be included in this sub-ontology.
- *Taxonomic Classification*. This sub-ontology would allow OntoBio to capture the taxonomic structure detailed for any specimen. This sub-ontology must follow the latest change of the botanical international nomenclature, which accepts *phylum* and *division* the as same taxonomic level.

3. OntoBio's Evolution Based on Tacit Knowledge Through a Conceptual Framework

In general, tacit knowledge modeling is not considered part of the formal scientific research life cycle, but it can inspire hypothesis to get a scientific view of knowledge. When modelled and made available, knowledge (implicit-explicit) becomes essential in the process of generating new knowledge. There are still open questions related to the representation, modeling, formalization and integration of tacit knowledge. A conceptual framework can be used to integrate specialists' mental models, aiming to map semantic components of attachable structures to formal ontologies. It also explores semantic annotation for dissemination and reuse. The framework aggregates semantic expressiveness to formal ontologies, and uses OntoBio, as the object of study. The framework guides the management of scientific tacit knowledge presenting different levels of representation, and allowing to retain knowledge to answer questions that OntoBio cannot currently respond.

The application of the conceptual framework to integrate scientific tacit knowledge applied to OntoBio [Albuquerque et al., 2016] suggested some

recommendations of change. These changes are associated to Mental Models (MMs) elicited and are:

- Create a *formal relation* (*can have*) between *Biotic Entity* (1..*) and *Popular Name* (1..*). This means that a *Biotic Entity* can be associated to multiple *Popular Names* and that a *Popular Name* can be associated to more than one *Biotic Entity* (MMs 1 to 10);
- To specialize *Macro Environment/Aquatic* into *Macrophyte Bank* (MM1), into *Soaked Trunk* (MM2), into *Well* and *Bench of Submerged Leaves* (MM5), into *Submerged Branches* (MM6), into *Inland Water Transition Zone* (MM9), into *Leaves Bunch* (MM10) and into *Mainland Igarapé*[2] (MM12);
- To specialize *Collection Method* into *Bait* (MMs 3, 13a, 13b) and into *Hand Net* (MMs 7a, 7b, 9);
- Create a *formal relation* (*feeds on*) between *Material Entity* (1..*) and *Biotic Entity* (1..*). This means that a *Biotic Entity* can eat multiple *Material Entities* and that a *Material Entity* can be the food of more than one *Biotic Entity* (MM3, 4, 13a, 13b);
- Create a *formal relation* between *Environment* (1..*) and *Collection Method* (1..*). This means that an *Environment* can be adopted to more than one *Collection Method* and that a *Collection Method* can be used in more than one *Environment*, depending on the Biotic Entity that is going to be collected (MM8);
- Create a *formal relation* between *Biotic Entity* (1..*) and *Habitat* (1..*). This means that a *Biotic Entity* can have multiple *Habitats* and that a *Habitat* can be used by more than one *Biotic Entity* (MMs 11a, 11b, 11c);
- Create a *component of* relation (*composed by*) between *Habitat* (1..*) and *Environment* (1..*). This means that a *Habitat* is composed by multiple *Environments* and that an *Environment* can be part of more than one *Habitat* (MMs 11a, 11b, 11c);
- Create a new concept *Organ* and instantiate it with *Flower* (MM14);
- Create a *formal relation* (*has*) between *Biotic Entity* and *Organ* (MM14);
- Create a *self formal relation* (*occurs*) between *Biotic Entity* (1..*) and *Biotic Entity* (1..*). This means that an organism's occurrence is subjected to the occurrence of another organism (MM14).

The original OntoBio and a trial version of OntoBio with some of these changes can be found at portal.inpa.gov.br/ctin/lis/ontobio/. More details of the framework and the formalization files used to apply it to generate OntoBio's recommendations for evolution can be found at portal.inpa.gov.br/ctin/lis/frameworkconceitual/.

[2] Small body of water, generally a tributary river or a canal. It's a word used by indigenous Tupi tribes when referring to a small strait or canal between two islands, or between an island and the mainland. *Igarapés* can only give way to small vessels (such as canoes, hence its Tupi denomination), as they are shallow, and ordinarily have very dark waters, being located deep within wealds or Amazonian thickets or forests.

4. Implementational Issues

OntoBio is developed using tools in a sequential order to provide a better code result. The ontological schema must be designed in a tool with graphical support to UML, such as Sparx Enterprise System Architect[3] (EA). EA is used to design OntoBio's ontological schema using OntoUML primitives. Once the ontological schema is concluded in a (.eap) file format, it can be exported to (.xmi) file format to be used in OntoUML Lightweight Editor (OLED)[4], current version named Menthor Editor[5].

The ontological schema in (.xmi) file format must be imported by Menthor and then can be converted to a (.owl) file format.

The final implementation phase is to use the (.owl) file in an OWL editor such as Protégé. The editor manipulates OWL ontologies and also provides a list of inference tools for testing the logical ontology consistencies.

Some implementation issues emerged with OntoBio's evaluation and are regarded as ontology modeling language limitations in the specific domain of biodiversity:

- OntoUML does not support Ø representation for anything (Ø..1, Ø..N). Limitation of cardinality representation. It justifies the adoption of a taxonomic representation with three ranks in original OntoBio – *family*, *genus*, *species*. All specie is associated to a *family*, a *genus*, a *specie*;
- OntoUML does not support the use of *high order*, essential for taxonomic classification. It supports only *kind* that does not model these concepts more appropriately;
- OntoUML does not allow modeling a sub-collection of a sub-collection. Ex.: States are sub-collections of countries; cities are sub-collections of states;
- There are inconsistences in the .owl file generated from the .eap file. Even if Menthor allows the automatic generation of OWL code of the ontological scheme designed, it is important to remember that a language in the level of analysis to design ontologies as OntoUML has more expressiveness power than a language for ontologies in the level of implementation, such as OWL. Thus, a code generated automatically in OWL does not reflect the reality modeled. Adjustments are required to maintain the integrity of that which has been patterned, thus justifying the use of Protégé. This is a recurring issue in the development of ontologies that still requires additional research and well elaborated solutions;
- OntoUML is based on UML and OWL is based on set theories. It implies that these ontology languages do not have a directly mapping between them. Some OntoUML definitions may be missing in OWL mapping at the Application level;
- Protégé does not support powertype that can be used in OntoUML.

[3] http://www.sparxsystems.com.au/products/ea/
[4] https://github.com/nemo-ufes/ontouml-lightweight-editor
[5] http://www.menthor.net/menthor-editor.html

5. Conclusions

This research revealed some conceptual misunderstandings and ontology language limitations. These issues must be dealt with according to the domain allowing the ontology to evolve in resources. Despite its limitations, OntoUML is a highly expressive formal ontology modeling language capable of guaranteeing less risk of semantic expressiveness loss than other ontology modeling languages. It produces logically and ontologically consistent models, but to do so, it is necessary to: 1) understand the meaning of each stereotype in OntoUML in order to use the appropriate meta-category for concepts; and 2) validate the ontology modelled by checking all the model's possibilities. A complex domain such as biodiversity, facilitates the identification of limitations in OntoUML and as a result, generates demands for improvements in the language. When these bottlenecks are solved, the ontology engineer will benefit with more resources for modeling.

Acknowledgement

We would like to thank GSI-ICOMP-UFAM, LIS-INPA, FAPEAM (Foundation for the State of Amazonas Research), Grant Number 021/2011 062.03101 / 2012-DO and CNPq (National Council for Scientific and Technological Development) Grant Number 486333 / 2011-6 for partially funding this research.

References

Albuquerque, A.C.F.; Santos, J.L.C.; Castro JR, A.N. (2015) "OntoBio: A Biodiversity Domain Ontology for Amazonian". Proceedings of 48th Hawaii International Conference on System Sciences. Kauai, Hawaii, January 5th – 8th. ISBN: 978-1-4799-7367-5

Albuquerque, A.C.F.; Santos, J.L.C.; Castro Jr, A.N. (2016) "A Conceptual Framework to Integrate Scientific Tacit Knowledge". To appear in Proceedings of SAI Intelligent Systems Conference, IntelliSys 2016. London, UK, September 21st – 22nd. IEEE.

Falbo, R. et al. (1998) "A Systematic Approach for Building Ontologies. In Artificial Intelligence". IBERAMIA'98 (Proceedings of the 6th Ibero-American Conference on AI, 1998), Coelho, H. (Ed.): LNCS 1484 (Lecture Notes in Artificial Intelligence), pp. 349-360, Springer-Verlag Berlin Heidelberg, Lisbon, Portugal.

Guizzardi, G. (2005) "Ontological Foundations for Structural Conceptual Models". PhD Thesis (CUM LAUDE), University of Twente, The Netherlands. Published as the same name book in Telematica Institut Fundamental Research. Series No. 15, ISBN 90-75176-81-3 ISSN 1388-1795; No. 015; CTIT PhD-thesis, ISSN 1381-3617; No. 05-74. Holanda.

Jackson, D. (2002) "Alloy: A Lightweight Object Modelling Notation". Transactions on Software Engineering and Methodology (TOSEM), New York, v. 11.

Sales, T. P. (2012) "Identificação de Padrões de Erro em Modelagem Conceitual Por Meio de Validação de Ontologias OntoUML Utilizando ALLOY". Universidade Federal do Espírito Santo.