

Research in Progress

Integration of ontologies: the domain of bioinformatics

DOI: 10.3395/reciis.v1i1.32en



Maria Luiza de Almeida Campos

PhD in Information Science
Department of Information Science – Federal Fluminense University (UFF)
UFF-IBICT Postgraduate Program in Information Science, Niterói, Brazil

Abstract

The growth in the use of distributed architectures, especially in the Web context, has contributed to make it possible to use previously isolated information in an integrated way. Ontologies play a fundamental role in this integration, facilitating the semantic interoperability of heterogeneous distributed systems. This article is an output of the research project entitled “Integrating Ontologies: the domain of Bioinformatics and the problem of terminological compatibility” which aims to develop guidelines which allow the development, use and integration of ontologies employed in the description and retrieval of Bioinformatics resources and services.

Keywords

Ontology integration, domain ontology, bioinformatics, theoretical base and methodology in ontology, language compatibility and integration

The purpose of this article is to present research currently underway by the author, with support from the CNPq (National Council of Scientific and Technological Development) for the period 2005-2008, which sets out to research mechanisms for the use, development and integration of ontologies in the domain of Bioinformatics. The specific focus is the field which includes research into Genomes and Transcriptomes, with the aim of supporting studies in this area being developed by institutions such as the Oswaldo Cruz Foundation (Fiocruz), contributing theoretical and methodological aspects of Information Science related to its key competencies in the development of documentary languages and knowledge organization.

Specifically, the aim is to identify theoretical and methodological approaches from within the theoretical bases of Information Science and Terminology, in particular from the field related to constructing and ensuring compatibility of languages, which can be employed in the development, use and integration of ontologies.

With regards to the integration of ontologies, in the first stage of this research project the aim is to identify the ontologies for the domain of Genomes and Transcriptomes which already exist internationally and nationally through actions which allow comparison of the different taxonomical models used for the representation of the domains and their relationships,

identifying the levels of semantic and linguistic compatibility, in order to present a proposal for terminological harmonization, revealing guidelines for the integration of ontologies.

In relation to the theoretical and methodological principles which can help in the elaboration of ontologies, the second stage of the research will focus on the identification, through the International Consortium, of the aspects of the project which are not covered by the different ontologies under analysis, through the use of the method for achieving language compatibility, aiming to take into consideration the Brazilian specificities. The thematic domains which are not covered will then be defined as an empirical space for the development of methodological proposals for their modeling through the identification of theoretical and methodological bases which have been studied and systematized for the elaboration of ontologies.

In the area of Information Retrieval Systems, the organization and retrieval of information have always been contingent on the associated technology. Nowadays, databases of all kinds have increased in number with the availability of information through networks and particularly the Web. The retrieval of information content does not yet take place in a satisfactory way, due to the lack of adequate access tools which allow, for example, control over terminology.

To guarantee this precision, there is an identified need for taxonomical and terminological tools for the semantic treatment of information contained in databases which would enable, amongst other processes, the integration of information as an aid to the development of research in knowledge domains.

Semantic tools such as ontologies must be constructed in a computerized medium, encompassing domains of study and research in the Portuguese language, to be used in digital and virtual libraries, in knowledge management systems, in enabling processes of information sharing between researchers, as an aid to search tools generally and, mainly, as an instrument for the improvement of the processing and retrieval of information on the Internet.

Ontology is a set of standard concepts where terms and definitions must be accepted by a community active in a certain domain, and whose goal is to enable multiple agents to share knowledge. An ontology is made up of terms, definitions and axioms relating to them. Ontologies represent a powerful way of inter-relating systems. They are mainly developed to structure knowledge bases or for use as semantic tools to support interoperability of information systems.

Nowadays, the growth in the use of distributed architectures, especially in the context of the Web, open interfaces for access to databases, mediator technologies and standard formats for data exchange, has helped to ensure that data which was originally isolated can be made available for use in an integrated way. Ontologies take on a fundamental role in this

integration, making possible the semantic interoperability of heterogeneous distributed systems. In this way, ontologies establish foundations of conceptual meanings without which the Semantic Web would not be possible, due to the heterogeneity of the concepts which are represented (JACOB, 2003).

Heterogeneity has been identified as one of the most significant problems, and one of the most difficult to solve. It involves interoperability and cooperation between multiple sources of information, reflecting syntactic, semantic and structural differences between systems.

Semantic heterogeneity is currently the main stumbling block for semantic interoperability, representing a huge challenge for the integration of information on the Web. To deal with this problem it is necessary to search for a language which is capable of representing knowledge and rules, as well as inferring new data.

This will be possible through the inter-relationships between specific domain ontologies, which are premised on the rational use of metadata for describing data in a homogenous way, the systematic use of ontologies, filling the gaps between heterogeneous sources of data, and the use of semantic associations, addressing the interoperability between domains (ADAMS, 2002).

Within the domain of the development of ontologies, approaches to their construction are specific and limited. The Computer Sciences literature has given priority first to ontologies as vocabularies for specific domains, without a theoretical base, then as a set of rules and theoretical contributions, without elements to guide the construction of vocabularies which allow the elaboration of logical definitions. In addition, the existence of two big problems in the methodology can be observed (FERNÁNDEZ-LOPEZ, 1999; GUARINO, 2002; JONES, 1998; SURE, 2002) for the ontological project: the lack of a systematic explanation of how and where the theoretical approaches will be used within the process of elaborating them; the lack of stages focusing on the integration and maintenance of ontologies in the method in the majority of methodologies.

We believe that the theoretical and methodological contribution of Information Science, incorporating elements of studies in Theory of Terminology (WUESTER, 1981), Concept Theory (DAHLBERG, 1978), Theory of Classification (RANGANATHAN, 1967) and Language Compatibility (NEVILLE, 1970; DAHLBERG, 1981; 1983) may produce proposals which can be effectively applied. In addition, these areas may benefit from being active in a quite applied area of the issue, escaping from the complexity of an excessively formal approach. We hope that this project may come to enable greater integration of information, creating mechanisms to consolidate the participation of Brazil in the international consortia which include research into Genomes and Transcriptomes.

Bioinformatics research in Brazil has been developing studies which aim to provide an

environment which can offer semantic information about scientific resources, such as data and programs in this area, and enable the use of these resources in a collective way by the relevant scientific community. One of the aspects covered by the description and retrieval of these resources and services is related to the development of a standardized and consensual language to facilitate the understanding of vocabularies, which are often interdisciplinary.

In the field of genomics, initiatives by the international scientific community in the past few years have led to an explosive growth in the amount of biological information generated each day (HGP, 2003). The initial concern was the creation and maintenance of databases for storing biological information. As the genomic databases are filled, and the genomes sequenced, the focus of the research begins to move to the mapping of genomes for the analysis of the vast range of information which results from the functional description of genes through Molecular Biology and Bioinformatics. Linking up the data obtained by the various research projects around the world about the relationships between enzymes, genes, chemical components, diseases, species, types of cells, organs, and so on becomes fundamental, aiming to respond to questions such as: What is the protein which this gene codifies? What is the function of this protein in this organism? Is this gene similar to that other gene present in a different organism? (MÉNDES, 2005).

It is therefore important to consider the relevance for Bioinformatics research of the management, description and organization of scientific resources in the digital medium. Specifically, in this area, these resources are not always available to biologists, and often they must resort to using proprietary programs, located in other institutions. Therefore, in order that these teams and/or institutions may share scientific resources among themselves, it is necessary to find a common way of describing and accessing these resources, to facilitate their searching and integration.

For this reason, the large quantity of data being accumulated in different databases around the world needs to be annotated and interpreted based on the genomic information already available. It is therefore necessary that the different projects which are interested in exchanging and integrating information describe their data in a way which enables the retrieval of information in a consistent way. Initiatives in the field of terminological treatment have been presented through repositories of ontologies.

The Open Biological Ontologies (OBO) collection is a repository of terminologies developed for shared use across different biological and medical domains. Among the most popular vocabularies which make up the OBO a highlight is the Gene Ontology (GO). The GO contains terms relating to three categories of the field, i.e. cellular components, biological processes and cellular functions.

In Brazil, specifically in the area of genomic scientific applications, there is a research project underway entitled “Comparing Genomes and Transcriptomes: a Bioinformatics consortium for the development of a Web platform and integrated databases”, currently funded by the CNPq and coordinated by Dr Alberto M. R. Dávila at Fiocruz. One of the main aims of this project is to provide an environment which can offer semantic information about scientific resources, such as data and programs, in the area of Bioinformatics and enable the use of these resources in a collective way by the relevant scientific community. It has been using the GO for the annotations in its database. At the international level, as mentioned above, the Gene Ontology is well known, and includes terms relating to biological processes, cellular components and molecular functions, independent of species, among other initiatives. However, until now there has been no national or international identification of ontologies developed with a specific conceptual cross-section, such as Genomes and Transcriptomes, to meet the needs of the groups coordinated by Fiocruz. Despite international efforts, the Gene Ontology seems not to include concept classifications which would fully meet the needs of the research under development in Brazil: in some cases it is necessary to investigate the existing harmonization between terms and their conceptual content.

The expected results of the research include: guidelines for the integration of ontologies, involving questions relating to conceptual definitions; guidelines for the development of ontologies; modeling of the domain of Genomes and Transcriptomes. In this last aspect, studies relating to the modeling of domains, a fundamental principle for the elaboration of taxonomies for ontologies, have already been a focus of investigation in our studies and research (CAMPOS, 2004).

The project involves researchers, professors and undergraduate and postgraduate students from renowned institutions in the areas of knowledge associated with its thematic focus, thereby guaranteeing the development of trained human resources in areas which are strategic for the treatment, integration and retrieval of information but so far little known in the country.

In this first stage of the project we aim to reach the following goals: 1. review of the literature in Information Science, Ontology and Terminology; 2. analysis and identification of principles for the integration and development of ontologies; 3. survey and analysis of ontologies in the domain of Genomes and Transcriptomes.

At the moment, we are concentrating on the following areas: 1. a literature review on the integration and compatibility of languages in the area of Information Science, Computer Science and Terminology. As a result of the survey carried out, we are developing a database, where the information is

being processed, enabling access by the group of researchers involved; 2. a survey of existing ontologies in the domain of Genomes and Transcriptomes, with the aim of mapping the areas and sub-areas within the domains and helping researches in the identification of production in these areas; 3. studies relating to conceptual definitions.

In the case of ontologies, the definitions allow the possibility of semantic compatibility, as they describe the semantic content of the linguistic sign (the term). This description makes it possible for intelligent agents to understand the meaning of the term and establish inferences about these meanings, since the definition is made up of characteristics of concepts, which are also concepts which relate to each other, making up the semantic understanding of the terms in question.

In this way, the definitions are of fundamental importance for the elaboration of consistent ontologies. However, it is true that existing ontologies are today lacking a defining standard for their elaboration. This is quite problematic given the question about the compatibility between languages which operate in cooperative databases, as is the case for the research being carried out in Bioinformatics. One of this study's theoretical and practical aspirations is to bring together different fields around the problem of the definition and compatibility of languages.

Until the 1960s definitions had a more conceptual and philosophical nature, and set out what should be thought about a concept. However, to meet the needs of science, which has a less philosophical and more operational nature, theoretical work was necessary which focused on the elaboration of operational definitions, which aimed to relate a given concept, as well as concepts which indicated the closest genre and the specific difference, to other concepts which attempted to define certain operations where this concept would be applied, or rather, to the concept which would be observed if certain operations were carried out.

On the other hand, the question of definitions has also been a field of study in the area of Terminology since the 1930s, through WUESTER (1981), whose goal was to study the terms in the area of specialist language, in other words, of science itself.

According to DAHLBERG (1983a), the concept of definition can be presented as:

“the equivalence between a *definiendum* (that which must be defined) and a *definiens* (how something should be defined) with the purpose of determining the understanding of the *definiendum* in any kind of communication.”

Based on this explanation, DAHLBERG presents three types of definitions: nominal, ostensive and conceptual. The nominal definition is that where the *definiendum* is a verbal expression and the *definiens* is a textual equivalence of this term, such as for example $A = B$. The ostensive definition is that where the *definiens* is

established by pointing to the reference named by the *definiendum*, that is, $C = A$. As for the conceptual definition, also known as the real definition, it happens when the *definiens* contains the necessary characteristics of a reference named by the *definiendum*, that is, $C = B$ of A .

In the case of the questions which cover the compatibility and the integration between ontologies, we are interested in the study of conceptual and nominal definitions, as they permit compatibility on the semantic and linguistic levels, respectively.

Based on what has been presented here, as an important contribution of the project, we also support the convergence of methods and techniques from the two areas which are fundamental for the development of practices relevant to ontologies, Information Science and Computer Science, as well as an empirical space of application, that is, the domain of Bioinformatics. In most of the projects on this topic, a specific bias can be seen towards one of these areas, without consideration of the important contributions which the other area could make. Based on the researchers' previous experience, there was a firm conviction of the importance of an integrated and multidisciplinary approach to the treatment of ontologies.

Bibliographic references

ADAMS, K. The semantic web: differentiating between taxonomies and ontologies. **On line**, v.26, n.4, p.20-23, Jul.-Aug., 2002.

CAMPOS, M. L. A. Modelização de domínios de conhecimento: uma investigação de princípios fundamentais. **Ciência da Informação**, Brasília, v.33, n.1, 2004.

DAHLBERG, I. A referent-oriented analytical concept theory of interconcept. **International Classification**, v.5, n.3, p.142-150, 1978.

DAHLBERG, I. Towards establishment of compatibility between indexing languages. **International Classification**, v.8, n.2, p.88-91, 1981.

DAHLBERG, I. Terminological definitions: characteristics and demands. In: **Problèmes de la définition et de la synonymie en terminologie**. Québec, GIRSTERM, 1983a. p.15-34.

DAHLBERG, I. Conceptual compatibility of ordering systems. **International Classification**, v.10, n.2, p.5-8, 1983b.

FERNÁNDEZ-LÓPEZ, M. Overview of methodologies for building ontologies. PROCEEDINGS OF THE IJCAI-99 WORKSHOP ON ONTOLOGIES AND PROBLEM-SOLVING METHODS (KRR5), Stockholm, Sweden, 2 Aug. 1999.

HGP. **Human Genome Program**, US. Department of Energy, Genomics and its impact on science and society: A 2003 Primer, 2003. Available at: <<http://>

www.ornl.gov/sci/techresources/HumanGenome/publicat/primer2001/index.sht>.

HAMMOND, W. Dimensions in compatibility. In: NEWMAN, S. M. (Ed.). **Information systems compatibility**. Washington: Spartan Books, 1965. p.7-17.

JACOB, E.K. Ontologies and the semantic web. **Bulletin of the American Society for Information Science and Technology**, Apr.-May. 2003.


MENDES, P.N. Uma abordagem para a construção e uso de ontologias no suporte à integração e análise de dados genômicos. 2005 Dissertação (Mestrado em Ciência da Computação) Instituto de Matemática/NCE da UFRJ, Rio de Janeiro.

NEVILLE, H.H. Feasibility study of a scheme for reconciling thesauri covering a common subject.

Journal of Documentation, n.26, v.4, p.313-36, Dec. 1970.

RANGANATHAN, S.R. **Prolegomena to library classification**. Bombay: Asia Publishing House, 1967.

SURE, Y.; STAAB, S.; STUDER, R. Methodology for development and employment of ontology based knowledge management applications. **SIGMOD Record**, v.31, n.4, p.18-23, 2002.

WUESTER, E. L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'Ontologie, l'Informatique et les Sciences des Choses. In: RONDEAU, G.; FELBER, F. (Org.). **Textes choisis de terminologie: I: fondements théoriques de la terminologie**. Québec: GIRSTERM, 1981. p.57-114. 

About the author

Maria Luiza de Almeida Campos

Maria Luiza de Almeida Campos has a Doctor degree in Information Science at the *Instituto Brasileiro em Informação Científica e Tecnológica - IBICT/UFRJ*, [Brazilian Institute in Scientific and Technological Information]. She is Associate Professor and Head of the Information Science Department at the *Universidade Federal Fluminense*, and of the Post-Graduation in Information Science Program (UFF/IBICT). She has activities in teaching and research in the area of Information Organization and Recovery, Taxonomy; Ontology, Thesaurus Construction. She was also Guest Professor of *strictu sensu* post-graduation courses at the Post-Graduation in Informatics Course in UFRJ (2002-2004) and at *latu sensu* courses at the Improvement Level (Index Elaboration Course, year 1998-2000/USU; Knowledge Management Course, year 1998/USU; Thesaurus Course, year 1994/UFF; Classification Theory Course, year 1990/Unirio), and at the Specialization Level (Archives Planning, Organization and Direction Course – The Information Management, year 1996, 2007). She has been a member of the *Comissão Nacional de Princípios Terminológicos da Associação Brasileira de Normas Técnicas-ABNT* since its foundation in 1992. The institution has the purpose of elaborating the National Terminology Norms. She is now developing the research "Integration of Ontologies: the domination of Bio-informatics and the terminological consistency problem" with a scholarship in productivity from CNPq. Besides that, she is the Coordinator of the Research Group registered in the CNPq "Ontology and Taxonomy, theoretical and methodological aspects". Currently, she is working in several institutions as a consultant on the elaboration of taxonomies, thesaurus and on index policies, such as *Finep*; *Casa de Rui Barbosa*; *Fiocruz*; *Sesc*; *Iphan*; *Central Globo de Produções* and *Petrobras*. She is the author of the book "Linguagens Documentárias: teorias que fundamentam sua elaboração" ["Documentary Languages: theories that set up the grounds for their elaboration"] and articles published in national and international journals.