Available online at www.sciencedirect.com
ScienceDirect
Web Semantics: Science, Services and Agents on the World Wide Web 4 (2006) 196–206
JOURNAL OF Web Semantics
www.elsevier.com/locate/websem

SAMBO—A system for aligning and merging biomedical ontologies☆

Patrick Lambrix*, He Tan

Department of Computer and Information Science, Linköpings Universitet, 581 83 Linköping, Sweden

Received 15 February 2006; accepted 1 May 2006

Abstract

Due to the recent explosion of the amount of on-line accessible biomedical data and tools, finding and retrieving the relevant information is not an easy task. The vision of a Semantic Web for life sciences alleviates these difficulties. A key technology for the Semantic Web is ontologies. In recent years many biomedical ontologies have been developed and many of these ontologies contain overlapping information. To be able to use multiple ontologies they have to be aligned or merged. In this paper we propose a framework for aligning and merging ontologies. Further, we developed a system for aligning and merging biomedical ontologies (SAMBO) based on this framework. The framework is also a first step towards a general framework that can be used for comparative evaluations of alignment strategies and their combinations. In this paper we evaluated different strategies and their combinations in terms of quality and processing time and compared SAMBO with two other systems.

Keywords: Ontologies; Alignment; Merging; Biomedical informatics

1. Introduction

Researchers in various areas, e.g. medicine, agriculture and environmental sciences, use biomedical data sources and tools to answer different research questions or to solve various tasks [3], for instance, in drug discovery or in research on the influence of environmental factors on human health and diseases. During recent years an enormous amount of biomedical data has been generated. These data are spread in a large number of autonomous data sources that are often publicly available on the Web. There are also numerous tools available on the Web. Due to this recent explosion of the amount of on-line accessible data and tools, finding the relevant sources and retrieving the relevant information is not an easy task. Further, often information from different sources needs to be integrated. The vision of a Semantic Web for life sciences alleviates these difficulties [38,19]. A key technology for the Semantic Web is ontologies. The Semantic Web can be seen as an extension of the current Web in which information is given a well-defined meaning by annotating Web content with ontology terms.

Intuitively, ontologies (e.g. [18,14]) can be seen as defining the basic terms and relations of a domain of interest, as well as the rules for combining these terms and relations. Ontologies are used for communication between people and organizations by providing a common terminology over a domain. They provide the basis for interoperability between systems. They can be used for making the content in information sources explicit and serve as an index to a repository of information. Further, they can be used as a basis for integration of information sources and as a query model for information sources. They also support a clear separation of domain knowledge from application-based knowledge as well as validation of data sources. The benefits of using ontologies include reuse, sharing and portability of knowledge across platforms, and improved documentation, maintenance and reliability. Overall, ontologies lead to a better understanding of a field and to more effective and efficient handling of information in that field. In the field of bioinformatics, for instance, the work on ontologies is recognized as essential in some of the grand challenges of genomics research [3] and there is much international research cooperation for the development of ontologies (e.g. the Gene Ontology (GO) [13] and Open Biomedical Ontologies (OBO) [32] efforts) and the use of ontologies for the Semantic Web (e.g. the EU Network of Excellence REWERSE Working Group A2 [38]).

Many ontologies have already been developed and many of these ontologies contain overlapping information. In Fig. 1, for

☆ The home page for SAMBO is http://www.ida.liu.se/~iislab/projects/SAMBO/.

* Corresponding author. Tel.: +46 13 28 2605.
E-mail addresses: patla@ida.liu.se (P. Lambrix), hetan@ida.liu.se (H. Tan).

1570-8268/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2006.05.003

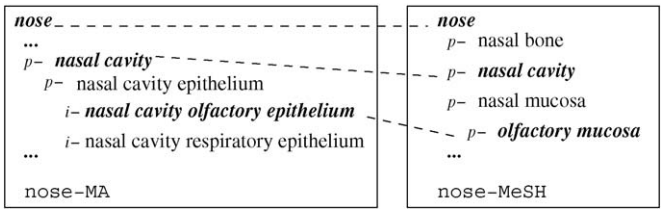

Fig. 1. Example of overlapping ontologies.

instance, we see two small pieces from two ontologies where terms in the two ontologies are equivalent (bold face). Often we would therefore want to be able to use multiple ontologies. For instance, companies may want to use community standard ontologies and use them together with company-specific ontologies. Applications may need to use ontologies from different areas or from different views on one area. Ontology builders may want to use already existing ontologies as the basis for the creation of new ontologies by extending the existing ontologies or by combining knowledge from different smaller ontologies. Further, different data sources in the same domain may have annotated their data with different but similar ontologies. In each of these cases it is important to know the relationships between the terms in the different ontologies. It has been realized that this is a major issue and some organizations have started to deal with it. For instance, the organization for Standards and Ontologies for Functional Genomics (SOFG) [42] developed the SOFG Anatomy Entry List which defines cross-species anatomical terms relevant to functional genomics and which can be used as an entry point to anatomical ontologies. In a similar spirit Ref. [41] defines a number of high-level relations in biomedical ontologies to promote interoperability of ontologies. In the remainder of this paper we say that we align two ontologies when we define the relationships between terms in the different ontologies. We merge two ontologies when we, based on the alignment relationships between the ontologies, create a new ontology containing the knowledge included in the source ontologies.

In this paper we tackle the problem of aligning and merging biomedical ontologies. Our contribution is three-fold: we present a framework for aligning and merging ontologies, develop an ontology alignment and merging system based on the framework and evaluate different alignment strategies and their combinations. The first contribution is presented in Section 3. We identified different types of alignment strategies and show how these strategies can be integrated in one framework. Most of the current alignment and merging systems can be seen as instantiations of our framework. Further, we developed a system for aligning and merging biomedical ontologies (SAMBO) according to this framework (Section 4). Within this system we have implemented some already existing alignment strategies as well as some new strategies. Although the framework and the SAMBO architecture are domain independent, we have focused on strategies that are applicable to the types of ontologies that are currently available in the biomedical domain.

We evaluated different alignment strategies and their combinations in terms of quality and processing time using several biomedical ontologies. We also compared SAMBO with two other systems. The results are discussed in Section 5. Related work is discussed in Section 6 and the paper concludes in Section 7. In the next section we provide some background on biomedical ontologies.

2. Biomedical ontologies

Ontologies differ regarding the kind of information they can represent. From a knowledge representation point of view ontologies can have the following components (e.g. [18,43]). Concepts represent sets or classes of entities in a domain. Instances represent the actual entities. Instances are, however, often not represented in ontologies. Further, there are many types of relations. Finally, axioms represent facts that are always true in the topic area of the ontology. These can be such things as domain restrictions, cardinality restrictions or disjointness restrictions. Depending on which of the components are represented and the kind of information that can be represented, we can distinguish between different kinds of ontologies such as controlled vocabularies, taxonomies, thesauri, data models, frame-based ontologies and knowledge-based ontologies. These different types of ontologies can be represented in a spectrum of representation formalisms ranging from very informal to strictly formal. For instance, some of the most expressive representation formalisms in use for ontologies are description logic-based languages such as OWL [34].

Biomedical ontologies (e.g. [18]) have been around for a while and their use has grown drastically since data source builders concerned with developing systems for different (model) organisms joined to create the Gene Ontology Consortium [13] in 1998. The research in biomedical ontologies is now also recognized as essential in some of the grand challenges of genomics research [3]. Further, the field has matured enough to develop standardization efforts. An example of this is the organization of the first conference on Standards and Ontologies for Functional Genomics in 2002 and the development of the SOFG resource on ontologies [42]. There exist ontologies that have reached the status of de facto standard and are being used extensively for annotation of data sources. Also, OBO was started as an umbrella web address for ontologies for use within the biomedical domain. Many biomedical ontologies are already available via OBO. There are also many overlapping ontologies available in the field. Most biomedical ontologies are vocabularies or taxonomies.

The ontologies that we use in our evaluations are GO ontologies, Signal-Ontology (SigO) [47], Medical Subject Headings (MeSH) [26] and the Anatomical Dictionary for the Adult Mouse (MA) [16]. The GO Consortium is a joint project whose goal is to produce a structured, precisely defined, common and dynamic controlled vocabulary that describes the roles of genes and proteins in all organisms. Currently, there are three independent ontologies publicly available over the Internet: biological process, molecular function and cellular component. The GO ontologies are a de facto standard and many different bio-data sources are today annotated with GO terms. The terms in GO are arranged as nodes in a directed acyclic graph, where multiple inheritances are allowed. The purpose of the SigO project