Ontological Analysis for the Unification of Biology

Domenico M. Pisanelli*, Massimo Battaglia^, Aldo Gangemi*, Geri Steve*
*** ISTC ^ INMM National Research Council, Roma, Italy**
pisanelli@acm.org

The overwhelming amount of information stored in various data repositories - including those available over the web - emphasizes the relevance of knowledge integration methodologies and techniques to facilitate data sharing. The need for such integration has been already perceived for several years, but telecommunications and networking are quickly and dramatically changing the scenario.

We are witnessing the unification of biology, since both biochemists and genetists now recognize a single universe of genes and proteins, and such unification is made possible also by the ever-increasing availability of the sequences of entire genomes.

However, such a growing demand has to rely on a solid conceptual foundation in order to give a precise semantics to the terabytes available in different genome databases and eventually traveling over the networks. The actual demand is not for a unique conceptualization, but for an unambiguous communication of complex and detailed concepts (possibly expressed in different languages), leaving each user free to make explicit his/her conceptualization.

Often this task is not an easy one to be achieved, since a deep analysis of the structure and the concepts of terminologies is needed. Such analyses can be performed by adopting an *ontological* approach for representing terminology systems and for integrating them in a set of ontologies.

Ontologies not only make knowledge re-use easier, they are also the foundation of standardization efforts since they make explicit the conceptualizations behind a terminology or a model. Therefore, an ontological analysis of a domain, which results into defining a library of ontologies, consist in the explicitation of implicit relationships among concepts in a given context.

We developed ONIONS, a methodology for ontological analysis. By means of this methodology we realized the library of ontologies ON9.2 (available at: http://saussure.irmkant.rm.cnr.it), including both general and domain specific ontologies [1]

We performed an ontological analysis of the Metathesaurus™ [2], a terminology data-bank developed in the context of the Unified Medical Language System (UMLS) project by the U.S. National Library of Medicine [3][4]. It collects millions of terms belonging to the most important nomenclatures and terminologies defined in the United States and in other countries too. About 500,000 preferred terms, named "concepts", have been chosen as representative of a set of synonyms and lexical variants in different languages. It is probably the largest repository of terminological knowledge in medicine.

Recently we started performing ontolological analysis in the genetics field.

As a case study, we investigated on the molecular function ontology defined by the Gene Ontology Consortium (http://www.geneontology.org) [5].

We implemented a wrapper translating from the XML ontology definition into LOOM, a formalism suitable for automatic classification [6].

The ontological analysis put in evidence the necessity of refining some assumptions made by the Gene Ontology developers. For example, metonymy is often used, since both enzymes and their functions are used in the same taxonomy.

Our experience has proved that the ontologies produced by means of the ONIONS methodology support:

- formal upgrading of terminology systems: term classification and definitions are now available in a common, expressive formal language;
- conceptual explicitness of terminology systems: (local) term definitions are now available, even though the source does not include them explicitly;
- ontological comparability, since pre-existing ontology libraries pertaining to different fields are largely employed.

In conclusion, we point out the following important features of ontologies:

- Semantic explicitness.
- An explicit taxonomy.
- Explicit linkage to concepts and relations from generic theories.
- Absence of polysemy within a given formal context.
- Modularity of contexts.
- Some *minimal* axiomatization to detail the difference among sibling concepts.
- A good naming policy.
- Rich documentation.

References

[1] Gangemi A, Pisanelli DM, Steve G, "An Overview of the ONIONS project: Applying Ontologies to the Integration of Medical Terminologies", *Data and Knowledge Engineering* , vol.31, pp. 183-220, 1999.

[2] Humphreys BL, Lindberg DA, "The Unified Medical Language System Project", *Proceedings of MEDINFO 92*, Amsterdam, Elsevier, 1992.

[3] Pisanelli DM, Gangemi A, Steve G, "An Ontological Analysis of the UMLS Metathesaurus", *JAMIA*, vol. 5 S4, pp. 810-814, 1998.

[4] Pisanelli DM, Gangemi A, Steve G, "A Medical Ontology Library that Integrates the UMLS Metathesaurus™", *Lecture Notes in Artificial Intelligence* 1620, Berlin, Springer Verlag, pp. 239-248, 1999.

[5] The Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology", *Nature Genetics*, vol.25, pp.25-29, 2000.

[6] MacGregor RM, "A Description Classifier for the Predicate Calculus" *Proceedings of AAAI 94, Conference*, 1994.