Ontology-based technologies — Technology transfer from bioinformatics?

Fabian Neuhaus, NIST

I. Introduction

In the call for paper for OIC 2008 the description of the conference contains the following optimistic outlook:

> New approaches are required to enable greater flexibility, precision, timeliness and automation of analysis in response to rapidly evolving threats. Ontology-based technology as applied in the areas such as bioinformatics has demonstrated the possibility of gains along all of these dimensions. The time is ripe to extend these gains to other spheres.

Ontology-based technologies clearly offer great potential for the intelligence community. In this paper I will discuss whether the intelligence community could adopt technologies that have been proven successful in bioinformatics. For this purpose we have to consider how biologists apply these technologies and how their needs differ from the needs of the intelligence community.

II. Kinds of knowledge

Biologists have been very successful at representing biological knowledge in a machine-readable form with the help of ontology-based technologies. However, we should not take for granted that the technologies that work for biologists would be appropriate for the intelligence community, because the kind of knowledge gathered by the intelligence community differs in important respects from biological knowledge. While the intelligence community is interested in individual people and organizations, biologists are producing scientific knowledge that consist of more or less general laws. Even in cases where biologists use terms from ontologies to describe the results of individual experiments, these results are formulated as laws; for example, laws like 'if a fruit fly has the mutation x, then the fly will have red eyes'. Biologists are only interested in the properties of individual animals or plants if these properties might provide evidence for or against a general hypothesis. For this reason, it is usually not necessary, and often not even possible, for biologists to keep track of the individual entities that they are experimenting with; e.g., no biologist would care to uniquely identify the individual fruit flies of a population, let alone the individual RNA molecules in a particular sample. In contrast, for the intelligence community it is crucial to identify individual persons of interest, to keep track of them over time, and to gather information about them. Furthermore, it is not the primary purpose of the intelligence community to produce and test general hypotheses.

III. Reasoning with instances

Most biological ontologies are written either in the OBO Flat File Format [1] or in OWL DL [2]. These ontologies are used primarily as controlled vocabularies; so far the use of biological ontologies for automatic reasoning has been surprisingly limited. However, even when biologists reason with the content of their ontologies, their needs typically differ from these in the intelligence community. Biologists are interested in type-level reasoning (so-called 'TBox reasoning'); the intelligence community is primarily interested in instance-level reasoning (so-called 'ABox reasoning'). For example, a biologist might be interested in the query 'What types of mutation lead to red eyes in fruit flies?' but a biologist would never enter the query 'Find all the fruit flies that have red eyes'. The reason is, of course, that biologists do not care about individual fruit flies; and they do not keep track of the individual animals.

In contrast, analysts in the intelligence community are interested primarily in instance-level queries about individual people and organizations and their properties and relations. For example, a typical query might be 'Find all people known to be member of Hamas, currently residents of Paris, and have been in Tehran in the last three years'. Since instance-level reasoning is irrelevant for biologists the OBO-format, which is the knowledge representation language that has been tailored to their needs, does not even allow assertions about instances. Consequently, all tools based on it do not support instance-level reasoning. Ontologies that are written in OWL DL can be used with reasoners like Pellet or Racer[1], which support instance-level reasoning. However, in spite of impressive performance improvements, as of 2008 these reasoners are not able to cope with the large-scale instance-level reasoning (ABox reasoning) that would be required by the intelligence community [3], [4], [5], [6].

IV. Time

Another difference between biological knowledge and the knowledge gathered by the intelligence community is related to time. Biological laws (and other natural laws) are timeless in the following sense: if a law like 'if a fruit fly has the mutation x, then the fly will have red eyes' is true then it is not only true now, but also at any given other time. Of course, this does not mean that biologists do not care about change over time. Evolutionary biology is strongly concerned with the changes of DNA that give rise to new species, and developmental biologists study the processes and changes that lead from fertilization to an adult organism. But while the individual organism changes over time during its development

[1] Any mention of commercial products or companies is for information only and does not imply recommendation or endorsement by the author or the National Institute of Standards and Technology.

(e.g., today's caterpillar is tomorrow's butterfly) the truth-value of statements about development in biological ontologies (e.g., 'The pupal stage follows the larval stage') does not change over time. As a result biologists have no need to express that a statement is true only with respect to a given time.

In contrast, much of the knowledge the intelligence community needs to represent is time-relative. For this reason, it turns out that the knowledge representation languages used by biologists do not meet the needs of the intelligence community. For example, it would be trivial to express a statement like 'All leaders of Hamas are located in the Gaza strip' in the OBO-format or in OWL DL but there is no straightforward way to express 'All leaders of Hamas are located in the Gaza strip on August 27, 2008.' The OBO-format cannot express statements about instances, but in OWL DL the same problem arises for statements about instances: e.g., there is no straightforward way to express 'John has been married to Sue in 2004 and John is married to Anne in 2008' in OWL DL.[2]

V. Sources

Biology, as any evolving science, contains competing theories that are inconsistent with each other. To maintain consistency, biologists limit the scope of their ontologies to textbook knowledge – knowledge that has been vetted by the community and is considered part of the scientific consensus. Obviously, this approach would not work for the intelligence community, which has to deal with conflicting information from unreliable sources. For this reason, it is crucial for the intelligence community to represent not only the information itself but also the sources of the information. A knowledge representation language suitable for the intelligence community would enable the representation of statements like 'Source x claims that Khaled Mashal will be in Tehran on August 17th or 19th'. One major advantage of representing sources and the information they provide within the same formalism is that the sources are treated as first-class citizens in the knowledge base and can be used in queries like: 'Are there two independent sources who claim that Khaled Mashai will be in Tehran?' or 'Provide source x and source y inconsistent information?'

The representation of and the reasoning about sources of information is far beyond the scope of the OBO-format as well as OWL DL. It is possible to stretch the boundaries of first-order logic in a way that one can represent information about sources. However, the resulting ontology is rather convoluted, and my experiments with Prover9 (a first-order logic reasoner [7]) showed that as a result the reasoner had difficulties to answer even fairly simple queries. A knowledge representation language that is designed to handle this kind of expression is the IKRIS Knowledge Language (IKL), an extension of the Common Logic Interchange Format [8], [9], [10]. Unfortunately, there are no reasoning engines for IKL available at this time.

[2]Note that it is possible to represent statements whose truth-values change over time in OWL DL, but the resulting ontologies are rather convoluted, and – at least in my opinion – OWL DL is a poor choice for ontologies that are intended to support reasoning with these kind of statements.

VI. Conclusion

There are some skills that biologists have developed when they adopted ontology-based technologies that might be relevant for the intelligence community: techniques to build and maintain large scale ontologies, evaluation methodologies, and general design principles for ontologies. However, biologists and the intelligence community deal with very different kinds of knowledge and create ontologies for different purposes. Thus the lessons that the intelligence community can learn from biologists will be limited: (i) The knowledge representation languages used by biologists do not meet the needs of the intelligence community. OWL DL is more expressive than the OBO-format, but since OWL DL offers no straight forward ways to deal with time-relative statements and offers no way to reason over the sources of statements OWL DL is still not expressive enough. (ii) Existing OWL DL reasoners are not able to handle the amount of instance-level reasoning that the intelligence community requires. (iii) Since the tools developed for biologists work with ontologies either in the OBO-format or in OWL DL it follows that these tools will not be useful for the work of the intelligence community.

References

[1] J. Day-Richter. The OBO Flat File Format specification, version 1.2. http://www.geneontology.org/GO.format.obo-1_2.shtml

[2] P.F. Patel-Schneider, P. Hayes, I. Horrocks. OWL Web Ontology Language semantics and abstract syntax. http://www.w3.org/TR/owl-semantics/

[3] Z. Pan Benchmarking DL reasoners using realistic ontologies. http://www.mindswap.org/2005/OWLWorkshop/sub6.pdf

[4] E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur, Y. Katz. Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol 5, issue 2, 2007, 51-53.

[5] Racer Systems. Release Notes for RacerPro 1.9.2 beta. http://www.sts.tu-harburg.de/%7Er.f.moeller/racer/Racer-1-9-2-beta-Release-Notes/release-notes-1-9-2.html

[6] J. Bock, P. Haase, Q. Ji, R. Volz. Benchmarking OWL reasoners. In F. van Harmelen, A. Herzig, P. Hitzler, Z. Lin, R. Piskac, G. Qi: *Proceedings of the Workshop on Advancing Reasoning on the Web: Scalability and Commonsense*, 2008.

[7] http://www.cs.unm.edu/ mccune/mace4

[8] P. Hayes, C. Menzel. IKL Specification Document. http://www.ihmc.us/users/phayes/IKL/SPEC/SPEC.html

[9] P. Hayes. IKL Guide. http://www.ihmc.us/users/phayes/IKL/GUIDE/GUIDE.html

[10] ISO/IEC 24707. Information technology – Common Logic (CL): a framework for a family of logic-based languages.