Towards more Challenging Problems for Ontology Matching Tools

Ernesto Jiménez-Ruiz* and Bernardo Cuenca Grau

Department of Computer Science, University of Oxford, UK
{ernesto,berg}@cs.ox.ac.uk

Abstract. We motivate the need for challenging problems in the evaluation of ontology matching tools. To address this need, we propose mapping sets between well-known biomedical ontologies that are based on the UMLS Metathesaurus. These mappings could be used as a basis for a new track in future OAEI campaigns.

1 Motivation and Background

The 2011 OAEI campaign consists of six different tracks. The so-called *Anatomy track* involves the largest test ontologies (containing between 2000-3000 classes).

Ontology matching tools have significantly improved in the last few years and there is a need for more challenging and realistic matching problems [1, 2] for which suitable "gold standards" exist.

There has been a long-standing interest within the bio-informatics research community in integrating thesauri, taxonomies and (more recently) also ontologies. The development of the UMLS-Metathesaurus (UMLS), which is currently the most comprehensive effort for integrating medical thesauri and ontologies, has been a very complex process combining automated techniques, expert assessment, and sophisticated auditing protocols [3–5].

2 Our Proposal

Although the standard UMLS distribution does not directly provide sets of "mappings" (in the OAEI sense) between the integrated ontologies, it is relatively straightforward to extract mapping sets from the information provided in the distribution files (e.g., see [6] for details).

Since UMLS-Meta integrates many widely used large-scale ontologies, such as FMA, NCI, SNOMED CT, or MeSH, we believe that the UMLS mappings between these ontologies could be used as a basis for a new track within the OAEI initiative. It has been noticed, however, that although these mappings have been manually curated by domain experts, they lead to a significant number of logical inconsistencies when integrated with the corresponding source ontologies (e.g., the integration of SNOMED CT and NCI via UMLS mappings leads to more than 20,000 unsatisfiable classes, as shown in Table 1).

* Ernesto Jimenez-Ruiz is supported by the EPSRC project LogMap

Ontologies	Original Mappings	Inconsistencies	Clean Mappings
FMA-NCI	3,024	655	2,898
FMA-SNOMED	9,072	6,179	8,111
SNOMED-NCI	19,622	20,944	18,322

Table 1. Repairing UMLS mappings (see [7])

To address this problem, we have presented in [6] and [7] several refinements of the UMLS mappings that do not lead to such inconsistencies. The mappings in [7] represent a larger subset of the UMLS-mappings than those in [6] as they were generated using "less aggressive" ontology repair techniques (see Table 1).

These "clean" subsets of UMLS mappings are readily available and could be used as reference alignments for a new, more challenging track within the OAEI (see `http://www.cs.ox.ac.uk/isg/projects/LogMap/`). In order to turn these reference alignments into a agreed-upon gold standard, some additional effort would be needed (e.g., manual curation). Another possibility would be to construct a "silver standard" by "harmonising" the UMLS mappings with the outputs of different matching tools over the relevant ontologies; similar silver standards have been developed for named entity recognition problems [8].

Although the use in an OAEI track of ontologies such as SNOMED CT, FMA and NCI represents a significant leap in complexity w.r.t. the existing anatomy track (from several million candidate mappings to several *billion*), we have recently developed a new matching tool, called LogMap [7], that is able to efficiently match these ontologies. We take our positive experiences with LogMap as an indication that a new track based on large-scale realistic ontologies and UMLS-mappings is not only feasible, but also potentially of great value, both for the developers of matching tools and the bio-informatics research community.

References

1. Shvaiko, P., Euzenat, J.: Ten challenges for ontology matching. In: On the Move to Meaningful Internet Systems (OTM Conferences). (2008)
2. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology Alignment Evaluation Initiative: six years of experience. J Data Semantics (2011)
3. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research **32** (2004)
4. Cimino, J.J., Min, H., Perl, Y.: Consistency across the hierarchies of the UMLS semantic network and metathesaurus. J of Biomedical Informatics **36**(6) (2003)
5. Geller, J., Perl, Y., Halper, M., Cornet, R.: Special issue on auditing of terminologies. Journal of Biomedical Informatics **42**(3) (2009) 407–411
6. Jiménez-Ruiz, E., Cuenca Grau, B., Horrocks, I., Berlanga, R.: Logic-based assessment of the compatibility of UMLS ontology sources. J Biomed. Sem. **2** (2011)
7. Jiménez-Ruiz, E., Cuenca Grau, B.: Logmap: Logic-based and scalable ontology matching. In: 10th International Semantic Web Conference (in press). (2011)
8. Rebholz-Schuhmann, D., et al.: CALBC Silver Standard Corpus. J Bioinform Comput Biol. (2010) 163–179