Towards a UMLS-based silver standard for matching biomedical ontologies

E. Jiménez-Ruiz[1], B. Cuenca Grau[2], I. Horrocks[2], and R. Berlanga[1]

[1] Universitat Jaume I, Spain, {ejimenez,berlanga}@uji.es
[2] University of Oxford, UK, {berg,ian.horrocks}@comlab.ox.ac.uk

Abstract. We propose a silver standard based on the UMLS Metathesaurus to align NCI, FMA and SNOMED CT. This silver standard aims at being exploited within the OAEI and SEALS Campaigns.

1 Motivation

The UMLS Metathesaurus (UMLS-Meta) [1] is currently the most comprehensive effort for integrating independently-developed medical thesauri and ontologies. UMLS-Meta is being used in many applications, including *PubMed* and *ClinicalTrials.gov*. The integration of new UMLS-Meta sources combines automatic techniques, expert assessment, and auditing protocols (see [2] for a review of current methods). In its 2009AA version, UMLS-Meta integrates more than one hundred thesauri and ontologies, including SNOMED CT, FMA and NCI, and contains more than 6 million entities. UMLS-Meta provides a list with more than two million unique identifiers (CUIs). Each CUI can be associated to entities belonging to different sources. Pairs of entities from different sources with the same CUI are synonyms and hence can be represented as an equivalence mapping. Thus, UMLS-Meta mappings could be considered as a *silver standard* to align ontologies such as SNOMED CT, FMA or NCI. The Ontology Alignment Evaluation Initiative (OAEI) could be benefited from UMLS-Meta so that it could be used as the input dataset for a new challenging track within the evaluation campaign. However, in our previous work [3, 4], we showed that UMLS-Meta contains a significant number of logic errors when the rich semantics of the ontology sources is taken into account together with the UMLS-Meta mappings.

2 Method and Results

Our experiments were based on the UMLS-Meta version 2009AA and the corresponding versions of FMA (version 2.0), NCI (version 08.05d) and SNOMED CT (version 20090131), which contain 66,724, 78,989 and 304,802 entities, respectively. After extracting the relevant parts of UMLS, we obtained 3,024 mapping axioms between FMA and NCI, 9,072 between FMA and SNOMED CT and 19,622 between SNOMED CT and NCI. Note that mappings are considered as OWL 2 axioms (see [5, 4]).

When reasoning over each of the source ontologies independently, all their entities were found satisfiable. However, after the respective integrations via UMLS-Meta mappings, we obtained a huge number of unsatisfiable entities, namely 5,015 when integrating FMA and NCI, 16,764 with FMA and SNOMED CT, and 76,025 with SNOMED CT and NCI.

We designed three logic-based principles [3, 4], namely *conservativity principle*, *consistency principle* and *locality principle*, to automatically detect and repair conflictive set of mappings. After the assessment, our automatic methods removed 570 (19%) of the mappings between FMA and NCI, 4,077 (45%) of those between FMA and SNOMED CT and 13,358 (63%) of those between SNOMED CT and NCI. When reasoning with the new revised mapping sets, we found only 2 unsatisfiable entities when integrating FMA and NCI, 44 for FMA and SNOMED CT, and none for SNOMED CT and NCI. These remaining errors were analyzed with our semi-automatic tool ContentMap [5] and they required to repair inherent incompatibilities between the source ontologies (see [4]).

3 Discussion

UMLS-Meta represents a reference of correspondences among ontologies, however, the direct integration of these ontologies with UMLS-Meta mappings leads to a huge number of unintended logical consequences. We propose instead a revised set of UMLS-Meta mappings which could be considered as a silver standard to evaluate ontology matching techniques over FMA, NCI and SNOMED CT. The proposed silver standard and related sources are available for download in `http://krono.act.uji.es/people/Ernesto/umlsassessment`.

The silver standard could be improved by ontology matching tools if they are able to find new valid correspondence which do not lead to logic errors. We also intend to design less aggressive techniques in order to preserve the maximum number of UMLS-Meta mappings. Additionally, ontology sources could be revised before the automatic assessment in order to smooth the incompatibilites.

References

[1] Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research **32**(Database issue) (2004) 267–270

[2] Geller, J., Perl, Y., Halper, M., Cornet, R.: Special issue on auditing of terminologies. Journal of Biomedical Informatics **42**(3) (2009) 407–411

[3] Jiménez-Ruiz, E., Cuenca Grau, B., Horrocks, I., Berlanga, R.: Towards a logic-based assessment of the compatibility of UMLS sources. In: Proceedings of the SWAT4LS workshop. Volume 559 of CEUR Workshop Proceedings. (2009)

[4] Jiménez-Ruiz, E., Cuenca Grau, B., Horrocks, I., Berlanga, R.: Logic-based assessment of the compatibility of UMLS ontology sources. Accepted for publication in BMC Journal of Biomedical Semantics (2010)

[5] Jiménez-Ruiz, E., Cuenca Grau, B., Horrocks, I., Berlanga, R.: Ontology integration using mappings: Towards getting the right logical consequences. In: Proc. of ESWC. Volume 5554 of LNCS., Springer-Verlag (2009) 173–187