Towards visualizing the alignment of large biomedical ontologies

Catia Pesquita[1], Daniel Faria[1], Emanuel Santos[1], Jean-Marc Neefs[2], and Francisco M. Couto[1]

[1] Dept. de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] Janssen Pharmaceutical Companies of Johnson & Johnson
cpesquita@di.fc.ul.pt

NOTICE: This is the author's version of a work accepted for publication. Changes resulting from the publishing process, including peer review, editing, corrections, structural formatting and other quality control mechanisms, may not be reflected in this document. Changes may have been made to this work since it was submitted for publication.

Abstract. To successfully integrate biomedical data it is crucial to establish meaningful relationships between the ontologies used to annotate this data. Recent developments in ontology alignment techniques, including our AgreementMakerLight system, have been successful in matching very large biomedical ontologies. However the visualization of these alignments is still a challenge.
We have developed a graphical user interface for AgreementMakerLight that follows its core focus on computational efficiency and the handling of very large ontologies. It allows non-expert users to easily align biomedical ontologies, offering a wide selection of matching strategies and algorithms, with a particular focus on the use of external background knowledge. The visualization of the resulting alignment is based on linked subgraphs which are generated according to search queries over the full graph composed by the matched ontologies and the mappings between them. This strategy decreases the need for computational resources and improves the visualization experience, by letting the user focus on selected areas of the alignment.

Keywords: Ontology Matching, Ontology Alignment, Alignment Visualization, Large Ontologies, Biomedical Ontologies

1 Introduction

Biomedical ontologies and controlled vocabularies are now a widely used technology to support the annotation of life sciences datasets. However, only by establishing meaningful connections across the concepts from various ontologies can we fully explore the knowledge they contain. Ontology matching techniques can accomplish this since they create mappings (i.e., correspondences) between semantically related entities belonging to different ontologies [1]. Ontology matching systems usually employ several ontology matching techniques both at the element and structural level which are then combined to produce a final alignment.
There are several challenges in matching biomedical ontologies, which arise from their characteristics. For instance, one of the main components of biomedical

ontologies is their textual information, in the form of labels, synonyms and definitions. Successful ontology matching systems need to be able to handle this richness, and also the inherent complexity of biomedical terminology. Furthermore, the domains covered by biomedical ontologies are frequently very large and detailed, with many biomedical ontologies possessing tens of thousands of classes dedicated to highly specific areas such as genomics, phenotypes or cellular structures. However, there are also opportunities within the biomedical domain such as the abundance of scientific literature or the availability of many related biomedical ontologies. Although there is a community effort to ensure orthogonality between ontologies as much as possible [2], there is still a significant overlap between many of them. In a recent visualization effort of the mappings between BioPortal [3] ontologies it has been shown that there are 254 ontologies with at least one mapping to another ontology. These mappings have been created through strict string matching and thus represent only a fraction of the true overlap between ontologies. In fact, at the time of writing this paper there were 373 ontologies in BioPortal and about 13 million mappings.

In order to address these issues, recent ontology matching systems have begun to include more elaborate strategies, such as creating highly efficient data structures or modularization approaches to handle very large ontologies [4, 5], tailoring of string similarity metrics [6] and exploration of different synonym types [7], ontology repair techniques to ensure the coherence of the alignments [5, 4], and the use of external resources and ontologies to increase the amount of available knowledge to support matching [5, 8].

An important feature of ontology matching systems is the ability to visualize the alignments between the ontologies, particularly in the biomedical domain where many of the end-users are not computer science experts. There are two main purposes in alignment visualization: supporting the navigation and inspection of mappings; and supporting interactive matching, whereby users can mark mappings as correct or incorrect, and even add new mappings [9, 4, 10]. These tasks are usually supported by two visual paradigms: trees and graphs [11]. Trees are particularly intuitive representations of hierarchical relations, however they are unable to represent multiple inheritance, and have to resort to duplication of classes, distorting the model. Graphs can handle both multiple inheritance and non-hierarchical relations, but can be less intuitive to use, particularly if the number of nodes shown is high. A recent evaluation of tree vs. graph based visualization has investigated the impact of individual ontology representation on the task of manual mapping evaluation [12]. In this study ontologies were represented either as trees or graphs and testers were given a list of mappings to evaluate. The results showed that trees are better suited to support list-checking activities, such as the evaluation of mappings, but graphs are more suitable to provide an overview, and thus better at supporting the creation of new mappings. Furthermore, for very large ontologies, with great depth and a large number of descendants per node, users struggle to preserve a mental model of the hierarchy when using trees, since the number of expandable nodes can be overwhelming. Graphs can partially circumvent this by allowing users to pan to areas of inter-

est, however visualization of a large number of nodes is also an issue.
However, ontology alignment visualization systems should consider not only how to represent the ontologies, but also the mappings between them. Furthermore, there are additional challenges posed by biomedical ontologies: (1) biomedical ontologies are typically large, sometimes with tens of thousands of classes; (2) many biomedical ontologies can have multiple inheritance or possess more than one kind of hierarchical relation (e.g., a taxonomy as well as a partonomy); and (3) non-hierarchical relations are also common, e.g. *regulates*, *has_substrate*, *has_role*, *participates_in*, etc. However, the matching of very large ontologies has only recently begun to be addressed by systems, and as a result current ontology matching systems with visualization capabilities are not well suited to either match or visualize very large ontologies with these characteristics.
In previous work we have developed a novel ontology matching system, AgreementMakerLight [5], derived from AgreementMaker, but specifically tailored to match very large ontologies. Here we present a graphical user interface for AgreementMakerLight, which supports the matching of large ontologies with several distinct parameters, including the use of background knowledge. The GUI also supports a graph-based visualization of mappings, that highlights the integration of both ontologies in a modular fashion.

2 Related Work

Most ontology alignment visualization systems display ontologies as trees, which the user can navigate, while mappings are shown as lines between the two ontologies [13, 14] or displayed in a table [15]. We have surveyed three freely and currently available ontology matching systems with visualization capabilities: AgreementMaker, COMA 3.0 and Optima.

AgreementMaker [13] represents ontologies as indented trees on side by side scroll-enabled panes. A mapping between two classes is represented by a straight line indicating the similarity score of the mapping. There is support for the visualization of several alignments over the same ontologies, using different colored lines for mappings of different alignments. When clicking on a node, users can see the properties of the corresponding class in a separate pane. However, AgreementMaker is unable to handle ontologies with tens of thousands of classes.
COMA 3.0 Community Edition [14] depicts ontologies as indented graphs in side-by-side scroll-enabled panels. When a node is clicked, the main label is shown along with the path to the root node in the form of coma separated labels. Mappings are colored according to their score. It is possible to compute different matching workflows over the same input ontologies, but you can only visualize one at a time. Different matching results can be merged or intersected, and their differences can be also be calculated. Furthermore, the tool is not optimized to handle large ontologies. Neither COMA 3.0 nor AgreementMaker allow for the visualization of non-hierarchical relations, nor of multiple inheritance.
Optima [16] displays each ontology as a graph in a window without zoom capabil-

ities, which severely limits its usability for large ontologies, since all nodes need to fit in a constrained area. Mapped nodes are highlighted, and when clicked, their label is shown and when double-clicked the matched node label in the other ontology appears. There is no graphical representation of mappings, nor any listing. Furthermore, the matching technique employed by Optima is also unsuitable to handle large ontologies.

3 AgreementMakerLight

3.1 Framework

The AgreementMakerLight (AML) is a lightweight framework for ontology matching based on the AgreementMaker system, which has been optimized to handle the matching of larger ontologies. Like AgreementMaker, the AML ontology matching module was designed with flexibility and extensibility in mind, and thus allows for the inclusion of virtually any matching algorithm. A key component of AML is the use of background knowledge sources which have been shown to improve the alignment of biomedical ontologies, as evidenced by AML achieving top results in several OAEI 2013 tracks [17].

3.2 Graphical User Interface

The graphical user interface of AML is divided in two areas: a Resource Panel where information about the ontologies and the alignment is shown (e.g.: number of classes, properties, mappings and performance metrics against a reference alignment), and a Mapping Viewer dedicated to the graph visualization of ontologies and mappings (Figure 1).
AML-GUI allows the user to load ontologies in OWL or RDFS and then opt between loading a precomputed alignment (encoded in RDF or as a simple tab-separated text file) and matching the ontologies. There are three pre-defined matchers to choose from: a simple Lexical Matcher, the AML matcher and the OAEI 2013 matcher. The Lexical Matcher is based on name and synonym string identity and is very efficient and generally precise. The AML matcher is an ensemble of string and lexical matching algorithms, with the option to choose several background knowledge sources to use in the matching process (see Figure 2). The OAEI 2013 matcher corresponds to the AML configuration used in OAEI 2013. All matchers have the option to set a cardinality for the alignment (strict one-to-one, permissive one-to-one and many-to-many), and also a threshold to select mappings to include in the final alignment. Both of the latter matchers have the option to perform a repair of the final alignment [18]. Finally, the user can also evaluate the produced alignment against a reference standard, and save it either in RDF or as a tab-separated text file. Once an alignment has been loaded or computed, the user can access a mapping in three different ways: by iterating over all mappings, via the next/previous mapping option; by selecting a mapping from the list of all mappings; or by querying the alignment for a search term contained in the name of a participant ontology class. This search is supported by an auto-complete function.

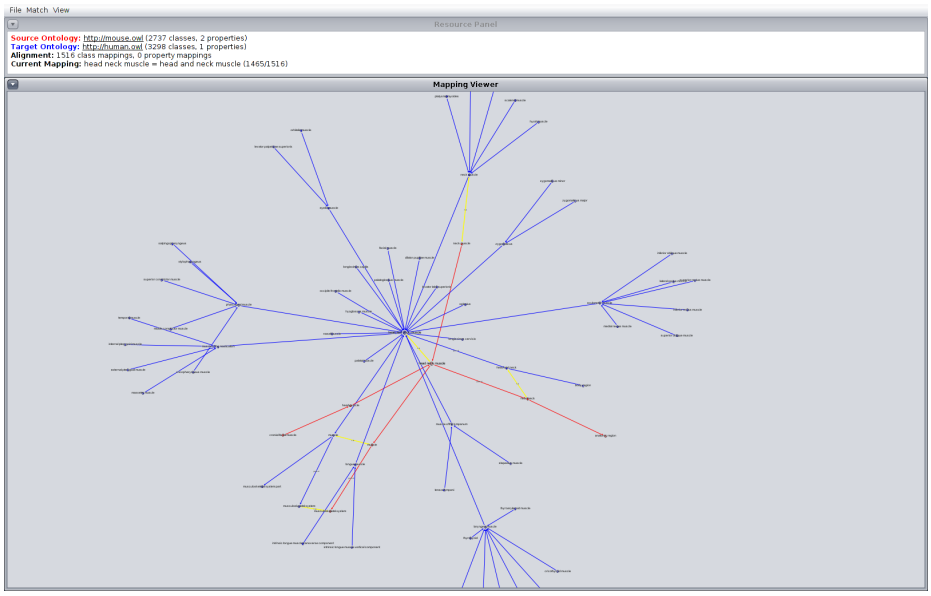

Fig. 1: Visualization of a mapping between anatomical ontologies in AML-GUI.

4 Visualizing Ontology Alignments

AML uses a graph to represent the mapped classes and their neighborhood, which is implemented using the Gephi API [19]. Once the user has selected a mapping to visualize, she can further specify the characteristics of its graph representation, by indicating whether the graph should show ancestor and descendant classes, and the distance between the classes involved in the selected mapping and their ancestors/descendants (from one to a maximum of five edges of distance). By default, AML shows both ancestors and descendants at a distance of two. Both ontologies are represented in the same graph, nodes and edges of the source ontology in red and of the target ontology in blue. Nodes are labeled with the classes main labels or names. Ontology edges are labeled with their relation type, except in the case of subsumption relations, which have no label. Directed edges are represented as arrows. Mappings are represented as yellow edges and labeled with their confidence score. Equivalence mappings are represented as double-edged arrows. All mappings between the ontology classes in the selected neighborhood are shown. The user can pan and zoom the graph, and at any time change the visualization options for the selected mapping, generating a new graph.

The following example focuses on the mapping between two classes of the Mouse and Human anatomy ontologies used in OAEI: “head/neck muscle” and “Head and Neck Muscle”. Figure 3 shows the representation of the mapping in AgreementMaker. Mapped classes are shown as colored nodes, and the mapping as a line between nodes. It is possible to see the direct descendants and ances-

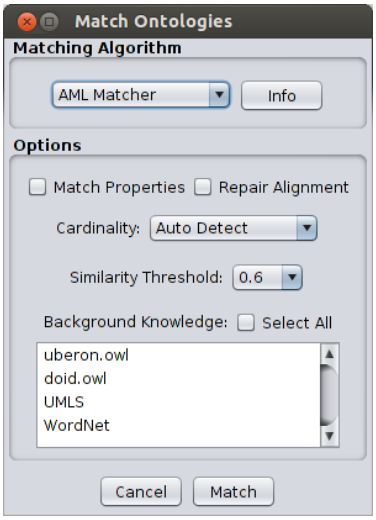

Fig. 2: Configuration window for the AML matcher.

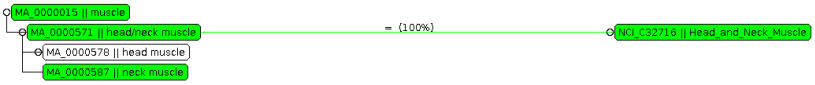

Fig. 3: Visualization of a mapping between anatomical ontologies in AgreementMaker.

tors of one of the mapped classes, which are also colored when they are mapped. However, it is not possible to see the neighborhood classes for both ontologies at the same time, and likewise it is not possible to see the mappings in this area. Figure 1 shows the same mapping in AML, with default settings. In the shown ontology subgraphs, there are four other mappings, both between ancestors and descendants of the selected classes. The graph representation allows the observation of several characteristics of the neighboring region of the mapping which are not apparent in the AgreementMaker visualization: the Human Anatomy ontology (in blue) contains a considerably larger number of classes in the neighborhood, half of the Mouse Anatomy classes can be mapped to a Human Anatomy class, and one of the mappings is established between classes that are part of the partonomy hierarchy (see Figure 4). This information can be valuable not only to evaluate the correctness of mappings but also to shed light on how regions around mapped classes are modelled.

5 Conclusions

Visualizing ontology alignments is a key feature to support user validation. In AML we focused on addressing the challenges in visualizing biomedical ontologies alignments, particularly the large size of the ontologies and the existence of

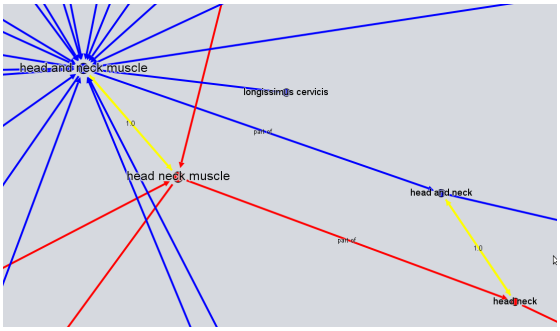

Fig. 4: Detail of a mapping between partonomy classes in anatomical ontologies in AML-GUI.

several types of relations between classes. Instead of allowing the visualization of full ontologies, which would be impractical in the case of very large ontologies, we have chosen to focus our visualization on the mappings. By selecting a particular mapping, users are shown a single graph composed of modules of both ontologies connected through their mappings. With this approach, we hope to better support the understanding of related areas within aligned ontologies, contrasting with the currently common approach of using linked trees in separate panes. Furthermore, by being graph-based, AML allows the visualization of several types of relations between ontology classes, including the cases of multiple-inheritance, which can be crucial to evaluate the validity of mappings. As future work, we plan to include dynamic graphs, graph color customization, and inspection of classes properties. AML is open-source and currently available both as a standalone executable jar file and as an Eclipse project at https://github.com/AgreementMakerLight/AML-Project.

Acknowledgements

CP, DF, ES and FMC were funded by the Portuguese FCT through the SOMER project (PTDC/EIA-EIA/119119/2010) and LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

References

1. Euzenat, J., Shvaiko, P.: Ontology matching. Volume 18. Springer Berlin (2007)
2. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology **25**(11) (2007) 1251–1255
3. Kocbek, S., Perret, J.L., Kim, J.D.: Visual presentation of mappings between biomedical ontologies. SWAT4LS (2012)

4. Jiménez-Ruiz, E., Grau, B.C., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. In: ECAI. Volume 242. (2012) 444–449
5. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M.: The agreementmakerlight ontology matching system. In: On the Move to Meaningful Internet Systems: OTM 2013 Conferences, Springer (2013) 527–541
6. Cheatham, M., Hitzler, P.: String similarity metrics for ontology alignment. In: The Semantic Web–ISWC 2013. Springer (2013) 294–309
7. Pesquita, C., Faria, D., Stroe, C., Santos, E., Cruz, I.F., Couto, F.M.: Whats in a nym? Synonyms in Biomedical Ontology Matching. In: The Semantic Web–ISWC 2013. Springer (2013) 526–541
8. Hartung, M., Gross, A., Kirsten, T., Rahm, E.: Effective mapping composition for biomedical ontologies. In: Semantic Interoperability in Medical Informatics (SIMI-12), Workshop at ESWC. Volume 12. (2012)
9. Paulheim, H., Hertling, S., Ritze, D.: Towards evaluating interactive ontology matching tools. In: The Semantic Web: Semantics and Big Data. Springer (2013) 31–45
10. Cruz, I.F., Stroe, C., Palmonari, M.: Interactive user feedback in ontology matching using signature vectors. In: Data Engineering (ICDE), 2012 IEEE 28th International Conference on, IEEE (2012) 1321–1324
11. Granitzer, M., Sabol, V., Onn, K.W., Lukose, D., Tochtermann, K.: Ontology alignmenta survey with focus on visually supported semi-automatic techniques. Future Internet **2**(3) (2010) 238–258
12. Fu, B., Noy, N.F., Storey, M.A.: Indented tree or graph? a usability study of ontology visualization techniques in the context of class mapping evaluation. In: The Semantic Web–ISWC 2013. Springer (2013) 117–134
13. Cruz, I.F., Sunna, W.: Structural Alignment Methods with Applications to Geospatial Ontologies. Transactions in GIS, Special Issue on Semantic Similarity Measurement and Geospatial Applications **12**(6) (2008) 683–711
14. Massmann, S., Raunich, S., Aumüller, D., Arnold, P., Rahm, E.: Evolution of the COMA match system. Ontology Matching (2011) 49
15. Ngo, D., Bellahsene, Z.: Yam++: a multi-strategy based approach for ontology matching task. In: Knowledge Engineering and Knowledge Management. Springer (2012) 421–425
16. Thayasivam, U., Doshi, P.: Optima results for oaei 2011. In: Proc. of 6th OM Workshop. (2011) 204–211
17. Grau, B.C., Dragisic, Z., Eckert, K., Euzenat, J., Ferrara, A., Granada, R., Ivanova, V., Jiménez-Ruiz, E., Kempf, A.O., Lambrix, P., et al.: Results of the ontology alignment evaluation initiative 2013. In: Proc. 8th ISWC workshop on ontology matching (OM). (2013) 61–100
18. Santos, E., Faria, D., Pesquita, C., Couto, F.: Ontology alignment repair through modularization and confidence-based heuristics. arXiv preprint arXiv:1307.5322 (2013)
19. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: ICWSM. (2009) 361–362