Methodology for Biomedical Ontology Matching

Jana Vataščinová[0000−0001−9656−5564]

University of Economics, nám. W. Churchilla 1938/4, 13067 Prague, Czech Republic
xvatj00@vse.cz

Abstract. This paper introduces a dissertation project in the field of ontology matching. Ontology matching plays an important role in integration of various systems or in connecting data which use different ontologies. One of the domains where different ontologies are used and where large amount of data is being continuously generated is biomedicine. The goal of the dissertation project is to propose a methodology for matching of biomedical ontologies. The existing projects do not address the topic as a whole. The related projects provide a methodology for ontology matching in general or talk about matching of biomedical ontologies (solely receiving a set of mappings). The thesis should define and address this gap. As biomedical ontologies can be quite specific, the need for a methodology for matching of these ontologies arises. To achieve the goal of the dissertation, the analytic steps come first - literature review of the field of matching of biomedical ontologies, use of the current methodologies and evaluation of their efficiency and their limitations, research of characteristics of biomedical ontologies, evaluation of different matching tools, and evaluation of combined results from multiple matching tools. Then, the formulation of the methodology steps will follow. After the formulation of the methodology, it needs to be evaluated. The evaluation should compare results from the matching process which follows the proposed methodology with the result from another methodology. At the end, some results of the analytic steps are drafted and current work to follow is described.

Keywords: Ontology · Ontology Matching · Biomedical Ontology.

1 Motivation

Biomedicine is a domain which plays an important role in our society. Biomedicine is a branch of medical science that applies biological and physiological principles to clinical practice. The branch applies especially to biology and physiology and it can relate to many other categories in health and biological related fields [1]. This domain captures an immense amount of data and new biomedical data is constantly being generated from different studies, experiments, etc. An example of where biomedical data is being used and generated are pharmaceutical companies. Without a good data organization (to which a computer can understand), much of the information can be lost.

Biomedicine is one of the main domains where *ontologies* are used and are of a great importance [2]. Biomedical ontologies are such ontologies which describe

part (or whole) of a biomedical domain. Ontologies can be different across data and their usage can be supported by *ontology matching*.

The content of many biomedical ontologies overlaps [3]. This points to the need to combine them or to the need of their matching. The ontology matching process creates relations between ontologies which can be of a great importance. For example, if an application works with ontology A, it can address data described by ontology B using the correspondences (mappings) between ontology A and ontology B. Biomedical ontologies and their matching can be a challenging task. In biomedicine, concepts can have many synonyms or their meaning can be artificially bordered. Identical concepts can have different names in different ontologies, and the other way around, concepts with corresponding names can have different meanings. The meaning of two concepts from different ontologies can partially correspond, etc.

Ontology matching is the search for identical concepts (connections) in different ontologies. The way a specific part of reality is captured can differ in different designs, therefore, ontologies that belong to the same domain can contain differences. The goal of ontology matching is to determine the similarity of concepts, properties and instances based on their name, structure or logical interpretation.[4]

A practical application can be as follows. There is a pharmaceutical company that has its own laboratories where different studies with drugs are carried out. To describe the data from the studies, the company uses its own ontology. There is a public platform for biomedical data and experiments with their results which uses different ontologies. It is of interest for the company to compare these results with their own results. This can mean comparing the given drugs results in the public experiments and in the internal studies. This would be achieved by query rewriting with the use of ontology alignment.

The question is, *how to carry out the matching process? Which ontology matching algorithm should be applied? Do I need to apply solely one? How to apply the algorithms efficiently?* When pharmaceutical companies start using ontologies, these ontologies can be created by their automatic generation from, for example, Excel sheets. Such ontologies can contain inconsistencies and might not follow best practices. These ontologies are likely to be replaced by new ontologies in the future and therefore, the ontology matching process needs to be repeatable. So, *overall, what strategy should be adopted to match ontologies for biomedicine?*

The goal of the dissertation project is to address why the current methodologies are insufficient and should be extended, and to propose a methodology for biomedical ontology matching.

2 State of the Art

There are many ways how the ontology matching task can be carried out (manual, automatic), many tools have been developed for the subject area (ten tools participated in the Large Biomedical Ontologies track of OAEI 2018 - see further

in this chapter) and even a general methodology has been proposed. However, we are not aware of any methodologies for biomedical ontology matching specifically. And yet, this area carries a great importance and a great need for data processing.

Considering the topic of matching of biomedical ontologies, there are three main sources to be followed. In the first source, there is a methodology for ontology matching proposed, the other sources are projects which contain matching of biomedical ontologies.

The first resource is the *methodology for ontology matching* by Jérôme Euzenat and Pavel Schvaiko [5]. The methodology consists of eight steps:

1. define the characteristics of the concrete problem to solve,
2. find if available and suitable alignments exist for the given problem,
3. select or build a matcher if necessary,
4. run the matcher,
5. evaluate the obtained alignment,
6. improve it by reiterating the matching process,
7. document and share the satisfying alignment,
8. process the alignment via a generator suitable for the given application task.

The second main resource are the *OAEI*[1] (Ontology Alignment Evaluation Initiative) campaigns and the reports from their events, such as [6]. OAEI is an international initiative for organizing evaluations of ontology matching systems, which has been active since 2004. It includes biomedical ontology-oriented tracks, which provide a valuable resource for matching of biomedical resources, especially in finding suitable matching systems. The tracks that are of interest are the *Large Biomedical Ontologies* track, then the Anatomy track and the Disease and Phenotype track [6].

The third main resource is the paper *Tackling the challenges of matching biomedical ontologies* by Daniel Faria et al [7]. The authors describe strategies employed by matching systems to tackle the challenges of matching biomedical ontologies and measures the impact of the challenges themselves on matching performance. The paper talks about the large size of the biomedical ontologies, about biomedical domain and its rich and complex vocabulary, or about different modeling views on the domain which can lead to the mappings to be logically irreconcilable due to conflicting restrictions.

Considering methodologies for matching of biomedical ontologies, [8] introduces a method for mapping for life science linked data using the mappings from BioPortal[2]. Another method for ontology matching is described in [9], where mappings are obtained from human contributions. These sources solely provide a way how to obtain a set of mappings, which would correspond to one step of the overall methodology.

Other resources for the subject area are projects describing specific matching tools [10, 11] (these are some of the tools that successfully participated in the

[1] http://oaei.ontologymatching.org/
[2] https://bioportal.bioontology.org/

OAEI Large Biomedical Ontologies track), technologies or projects [12] (project for biomedical data collecting). To complement the methodology for ontology matching mentioned above, another description of the methodology by Euzenat can be found in [13].

3 Problem Statement and Contributions

The main goal of my dissertation is to propose an efficient methodology for matching of biomedical ontologies. This methodology should include the whole process of ontology matching, starting with identifying ontologies and characterizing needs and ending with the alignments implementation. The methodology will be based on the already existing methodology for ontology matching (see [5, 13]), and it will be specialized for matching of biomedical ontologies (reflecting their needs and characteristics).

The methodology for ontology matching by Euzenat and colleagues [5, 13] proposes only a general overview of the ontology matching process. Biomedical ontologies can be very specific and for their efficient matching, a more detailed methodology should be provided. Methodology for biomedical ontology matching should capture the common problems and characteristics for each step of the matching process. In this way, all the important steps should be taken into consideration when carrying out an biomedical ontology matching process.

One of the main questions that needs to be solved is: *What are the characteristics of biomedical ontologies and how do they influence the ontology matching process?* Biomedical ontologies can be quite specific with their representation and biomedical terms can have unique relations and characteristics. These characteristics can play a crucial role in the ontology matching process. For example, they can influence the choice of matching tools that will be used.

4 Research Methodology and Approach

In order to achieve this goal, the following methodology will be used:

- literature review of the field of matching of biomedical ontologies,
- use of the current methodologies and evaluation of their efficiency and their limitations,
- research of characteristics of biomedical ontologies,
- evaluation of different matching tools, and evaluation of combined results from multiple matching tools,
- methodology development,
- methodology evaluation.

Regarding the literature review, the goal is to gather all available information about matching of biomedical ontologies, about methodologies for ontology matching in general, and about biomedical ontologies. It is necessary to note all the known challenges and obstacles, as well as the solutions and best practices.

This knowledge should then be applied specifically to the matching of biomedical ontologies in cases when it applies to ontology matching in general or to biomedical ontologies.

The use of the current methodologies and their evaluation should result in the analysis of their limitation, which should be further addressed.

In order to obtain characteristics of biomedical ontologies, the first step is to get an overview of all the biomedical ontologies that are available and to select those that are the most 'important' (for example, those that are widely used or those representing more specific domains). For this purpose, the repository of biomedical ontologies BioPortal and Ontology Lookup Service[3] will be used. These ontologies are then to be reviewed with the goal of registering their potential characteristics. In order to receive statistically comparable characteristics, Online Ontology Set Picker[4] (OOSP) tool for obtaining statistics for ontologies may be used. The characteristics obtained are then, again, to be applied in relation to the matching process.

For the part of evaluation of different matching tools, the OAEI campaigns will be used as the main source. Specifically, the tracks which consider biomedical ontologies - mainly the Large Biomedical Ontologies track. Other tracks might be the Anatomy track and the Disease and Phenotype track. These tracks provide a summary of results of different ontology matching tools received for given pairs of ontologies. Thus, different ontology matching tools can be compared based on their results. All retrieved results of each tool will be examined with the goal of finding regularity. For example, *does one tool always reliably find mappings of some kind? If one tool finds mappings that none of the other tools found, are those mapping characteristic in some way? Those mapping that a tool returned that are not correct, is there any common characteristic of such mappings - or in other words, in what kind of cases is the tool making mistakes?* In case any common characteristics are found for the tools, these findings can be applied to improve the ontology matching process. Another step to be done is to analyze the mappings in the reference alignment which were not found by any of the tools. Results should then be used for improving the tools so that these matches will be found. Given the characteristics of the purpose of the ontology matching process and of the biomedical ontologies that are to be matched, suitable tools can be effectively chosen or combined.

For all of the findings, it is necessary to evaluate which one of the tools are applicable for biomedical ontologies.

After gathering all the information from the analysis above, the design of the methodology will follow. Here, the steps of the methodology will be proposed.

5 Evaluation Plan

The proposed methodology for biomedical ontology matching should be evaluated in the following ways. When carrying out a process of biomedical matching,

[3] https://www.ebi.ac.uk/ols/index
[4] https://owl.vse.cz/OOSP/

the process should provide better results when following the biomedical ontology matching methodology compared to when it follows a methodology for ontology matching that is not specifically created for matching of biomedical ontologies. The results can be presented to domain experts for evaluation.

For evaluating the analytic findings that should be used for the methodology forming, the existence of reference alignment is quite desirable due to the large size of the biomedical ontologies. Such an opportunity is provided by the existing challenges reflected by biomedical tracks within OAEI campaigns. These tracks provide reference alignments for evaluating the mappings returned by the tools.

This part should help with the evaluation of the correctness of the obtained alignment. However, the process of matching of biomedical ontologies should not end here. The process of the matching end with the final implementation of the alignment and its final use. The usability and effectiveness of the implemented alignment needs to be evaluated as well with regard to the purpose of the ontology matching process. For example, it might not be desirable for all the matching processes to consider only those mappings that are logically compatible.

6 Preliminary Results

As the dissertation is still in a very early stage, only some suggestions can be presented so far. These are some basic characteristics of biomedical ontologies (observed without using any tools yet), a connection between the purpose of the ontology matching process and the approach to receiving a set of mappings, and some observations from an (ongoing) analysis of the results of the Large Biomedical track of the OAEI campaigns 2018[5].

6.1 Biomedical Ontologies' Characteristics

As already mentioned, one of the important characteristics of biomedical ontologies is undeniably their large size compared to ontologies in the Linked Open Vocabularies[6]. In many cases, there can be tens or hundreds of thousand classes solely. This aspect plays an essential role in deciding between manual or automatic ontology matching - considering such size, it is impossible to use manual matching. In case of many mapping found by automatic matching, it might not be possible to even review the proposed mappings. It is, therefore, necessary to choose a reliable tool for automatic ontology matching and to know the risks we might choose to accept.

Another characteristic of biomedical ontologies is the naming of the concepts. In the biomedical ontologies, it is common to use different codes as the concepts' names. Therefore, it is necessary to work with the concepts' labels in the matching task. For example, one of the National Cancer Institute Thesaurus[7]

[5] oaei.ontologymatching.org/2018/

[6] https://lov.linkeddata.es/dataset/lov

[7] https://cbiit.cancer.gov/ncip/biomedical-informatics-resources/interoperability-and-semantics/terminology/

(NCIt) concepts is ncit:C16403 and its label is Cell Line. In the Semanticscience Integrated Ontology[8] (SIO), the concept with the same label is sio:010054.

In many cases, biomedical ontologies are also represented rather as taxonomies. Looking at the Large Biomedical Ontologies track of the OAEI campaigns, all three of the used ontologies have a taxonomic character (one of them being NCIT). In many cases, they do not define many properties. For example, NCIT defines 145,810 classes and only 97 properties. The hierarchical assignment for biomedical term is often artificial and not clear, thus it can be very different in various ontologies. The ontology matching tools are therefore left with mainly lexical matching.

6.2 Purpose of the matching of biomedical ontologies

One of the important aspects that should not be forgotten is the purpose of the process of matching of biomedical ontologies. The purpose should be clear and it should be among the first questions to be decided.

The purpose shall be reflected in the selection of automatic matching tool. For example, *is it desirable to use a tool which excludes mappings with a logical incoherence?* Next, looking back at an example of biomedical ontology use case in Section 1, the task of the matching process was query rewriting - seeing the given drugs results in the public experiments and in the internal studies. In this case, the goal is to find the corresponding drug without the need of the two classes from different ontologies to be interchangeable.

6.3 OAEI Large Biomedical Track 2018

The Large Biomedical track from the OAEI campaigns poses the tasks of matching three large biomedical ontologies (of hundreds of thousand classes). As seen from the results (those evaluated so far), the mappings were created largely based on the names of the concept containing the same word or synonyms. This corresponds with the fact that the selected ontologies have taxonomy characters with differently constructed hierarchies. These ontologies also have very few properties compared to their size (all having less than 200 properties).

7 Conclusions

Creating a methodology for biomedical ontology matching is a challenging task. It should be of use to those workplaces where biomedical ontology matching is needed and where it can enable data sharing, data retrieval, etc.

The immediate work to follow inside the analysis of the OAEI Large Biomedical track is to analyze the mappings in the reference alignment which were not found by any of the tools or to analyze the mappings discovered by the tools that were not included in the reference mappings. Both tasks also point to the need

[8] http://sio.semanticscience.org/

to investigate the source of the reference alignment and its reliability. Looking at the mappings discovered by the tools that were not included in the reference mappings so far, most of the mapping included the same word or its synonym, but not in the equivalent meaning. However, some mappings that appear to be correct that were not included in the reference mappings were found as well (having identical name or different names with the same meaning).

Acknowledgement

The research has been partially supported by IGA VSE 33/2019. Additionally, I would like to thank my PhD supervisor Vojtěch Svátek, my PhD adviser Ondřej Zamazal and all the ESWC 2019 reviewers for their valuable comments and feedback.

References

1. Biomedicine, http://www.memidex.com/biomedicine. Last accessed 10 February 2019
2. Rubin. D.L., Shah, N.H., Noy, N.F.: Biomedical ontologies: a functional perspective. Briefings in Bioinformatics **9**(1), 75–90 (2008)
3. Gross, A., Hartung, M., Kirsten, T., Rahm, E.: Mapping Composition for MatchingLarge Life Science Ontologies. In: Proceedings of the Second International Conference on Biomedical Ontology, pp. 109–116. Buffalo, NY (2011)
4. Staab, S., Studer, R.: Handbook on Ontologies. 2nd edn. Springer, Berlin (2009)
5. Euzenat, J., Schvaiko, P.: Ontology Matching. 2nd edn. Springer, Berlin (2013)
6. Algergawy, A., Cheatham, M., Faria, D., et al.: Results of the Ontology Alignment Evaluation Initiative 2018. In: Ontology Matching OM2018, pp. 76–116. CEUR-WS, Cáchy (2018)
7. Faria, D., Pesquita, C., Mott, I., Martins, C., Couto, F.M.: Tackling the Challenges of Matching Biomedical Ontologies. Journal of Biomedical Semantics **9**(4) (2018)
8. Zaveri, A., Dumontier, M.: Ontology Mapping for Life Science Linked Data. In: Proceedings of the First International Workshop on Biomedical Data Integration and Discovery. CEUR-WS, Japan (2016)
9. Sarasua, C., Simperl, E., Noy, N.: CROWDMAP: Crowdsourcing Ontology Alignment with Microtasks. Lecture Notes in Computer Science **7649**, 525541 (2012)
10. Jiménez-Ruiz, E., Grau, B.C., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. Proceedings of the 20th European Conference on Artificial Intelligence, 444–449 (2012)
11. Faria, D., Pesquita, C., Santos, E. et al.: The AgreementMakerLight Ontology Matching System. OTM Confederated International Conferences On the Move to Meaningful Internet Systems, 527–541 (2013)
12. PubChem, http://pubchemdocs.ncbi.nlm.nih.gov/rdf. Last accessed 10 February 2019
13. Euzenat, J., Le Duc, C.: Methodological guidelines for matching ontologies. In: Suárez Figueroa, M., Gómez Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology Engineering in a Networked World, pp. 257–278. Springer, Heidelberg (2012)