Holistic Multiple Ontologies Merging

Samira Babalou
(Early Stage PhD)

Heinz-Nixdorf Chair for Distributed Information Systems
Institute for Computer Science, Friedrich Schiller University Jena, Germany
samira.babalou@uni-jena.de

Abstract. Ontologies as the main infrastructure to represent data in the Semantic Web are widely developed independently in each area. When it comes to finding a suitable ontology for a given application, two problems occur: Often, an ontology will cover just a part of the domain of interest or competing ontologies modeling the domain from different viewpoints exist. Thus, before being able to leverage the power of ontologies, they themselves need to be integrated. This is a challenging task. The existing approaches are mostly limited to a binary merge. However, by the large availability of the relevant ontologies in the desired domain, an efficient multiple ontologies merging technique is often a necessity to overcome the scalability problem. This research thus advocates for the development of a holistic, efficient multiple ontologies merging method called $\mathcal{C}o\mathcal{M}erger$, to satisfy the scalability issue. For efficient processing, rather than merging a large number of ontologies, we merge a small number of clusters. To approve the feasibility of our approach, we will run $\mathcal{C}o\mathcal{M}erger$ on real-life datasets. Further, our platform will be freely accessible through a live portal.

Keywords: Semantic web . Ontology merging . Ontology mapping

1 Problem Statement

Ontologies are the semantic model to represent data on the Semantic Web. Often a domain has more than one "standard" ontology for the same general concepts. They either cover just a part of the domain of interest or model the domain from different viewpoints. In this fashion, multiple heterogeneous ontologies are independently developed in each domain. In real-world applications of the Semantic Web, this is an essential demand to interoperate with more than two ontologies toward acquiring the desired knowledge for scientists. Indeed, different ontologies cover particular aspects of a domain of discourse but overlap to a certain degree. Therefore, an efficient technique for merging multiple ontologies is often a necessity both during ontology development and when ontologies are used in conjunction with data at the query processing level. This can be a cost-efficient approach and saves a lot of development effort. Thus, before being able to leverage the power of ontologies, they themselves need to be integrated. This is a challenging task.

Existing ontology merging approaches [14,15,18,23] are mostly limited to merging only two ontologies, partly due to using a binary merge (i.e., merging two ontologies at a time). In principle, a series of binary merges can be applied to more than two ontologies, however, they are no longer sufficiently scalable and viable for a large number of ontologies [17]. Precisely, to merge n ontologies, a binary-merge approach needs to run the $\frac{n.(n-1)}{2}$ pairwise alignment processes and $(n-1)$ combination operations in an incremental fashion. Nevertheless, merging multiple ontologies ($n > 2$) at the same time has not been extensively studied mainly due to the much more complex search space and it still remains one of the key challenges in the future research agenda. Therefore, to overcome the binary merge limitations, the holistic strategy has been introduced as a feasible and efficient method in [17]. Following this approach, in our proposed framework $\mathcal{CoMerger}$, we advocate for developing an efficient, holistic merging technique that scales to many ontologies. It gets as an input a set of ontologies with their alignments and automatically generates a merged ontology with a set of output mappings between the merged and the input ontologies. At first, the n input ontologies will be clustered into k clusters. Afterward, the clusters will be combined based on the corresponding pairs to produce the merged ontology. To this end, the main problem statement of our research is to enable a holistic ontology merging method by extending Semantic Web technologies. Thus, the general research question that my thesis tries to address is:

Research Question: How can the holistic approach be applied for merging n ontologies to overcome the scalability problem?

To contribute to the main research question, two RQs with further sub-questions can be concluded.

RQ1- *How can the elements of input ontologies be effectively clustered into k clusters based on the detected correspondences to facilitate the merge process?*

> ***RQ1-1-*** *Does the clustering process lead to a significant reduction in execution time and complexity of merging process without compromising quality?*
> ***RQ1-2-*** *What is the effect of varying the number of clusters k on the merge quality?*

RQ2- *How can the combination of k clusters efficiently generate the merged ontology?*

> ***RQ2-1-*** *Which requirements should be considered for the merging step?*
> ***RQ2-2-*** *How to fulfill these requirements?*
> ***RQ2-3-*** *How can a high quality of the merge result be achieved?*
> ***RQ2-4-*** *How to accomplish consistency in the merged ontology?*

2 State of the Art

Merging strategies basically have been divided into two main categories [6]: *"binary"* and *"n-ary"*. The *binary* approach allows merging of two ontologies

at a time, while *n-ary* strategy lets to merge n ontologies ($n > 2$). To deal with merging more than two ontologies, the *binary* strategy needs the quadratic complexity of merging operations and also needs a final analysis to add missing global properties [6]. However, in the *n-ary* strategy, the number of merging steps is minimized. Moreover, a considerable amount of semantic analysis can be performed before merging, thus avoiding the necessity of a further analysis and transformation of the merged ontology. This approach also is called *"holistic"* strategy [17]. Indeed, with continuously increasing amount of data being produced, developing solutions to deal with the simultaneous merging of different ontologies is becoming necessary.

To process multiple ontologies, for instance, in the multiple ontologies matching scenario in [11], to match 4000 web-extracted ontologies on six computers using a pairwise strategy took about one year, which indicates the insufficient scalability of pairwise strategies. As a further example of multiple ontologies merging, the integration process in the biomedical ontology UMLS Metathesaurus [7] was highly complex and involved a significant effort by domain experts. To the best of our knowledge, the *holistic* ontology merging has not been practically applied and still is one of the key challenges in this field. As an example, Porsche [20] semi-automatically merges many tree-structured XML schemas and holistically clusters all matching elements in the nodes of the merged schema. The final merge result depends on the order in which the source schemas are matched and merged. Low alignment accuracy and low minimality on the merge result arise from its simple heuristic functions. Furthermore, consistency issues have not been considered.

Principally, general processes in the existing approaches seem to indicate two different strategies: *"one-level merge"* and *"two-levels merge"*. In the latter one, an intermediate merge result is produced at the first level. Then, in the second phase, the intermediate result is refined to produce a final merge result. In contrast, *one-level merge* tends to produce the merge result in one incrementally processing step [10,13]. In each element combination, they analyzed whether it does not have any inconsistencies with other previous merged elements. Although the *two-levels merge* is the most used strategy in the literature reviews [14,18,23], there is no comparison between the effectiveness of these two strategies. We have prioritized to use *one-level merge* strategy and check the arising inconsistencies before applying each combination.

3 Proposed Approach

This research aims to comprehensively address the holistic multiple ontologies merging issue to improve the shortcomings of previous approaches. We have been investigating the existing methods and found the scalability issue as an ongoing challenge. Therefore, we have four main objectives:

- **Overcoming the scalability problem in merging multiple ontologies:** We aim to develop a framework for holistic multiple ontologies merging to overcome the scalability issue.
- **Achieving a high accuracy in the merge result:** In order to gain a high accuracy, the merge requirements should be fulfilled, however, it should be possible to customize them to the task at hand. We plan to derive a method to apply these customized requirements in a way that they do not contradict each other.
- **Developing a web tool:** As a proof of concept, this research aims to develop a web tool for merging multiple ontologies. This tool can be divided into two sections: *merger* and *evaluator*. The latter includes systematic criteria for evaluating the merged ontology independent from the merge techniques.
- **Validating:** To evaluate our framework, we will carry out a set of experimental tests to analyze the performance of the tool and the merge algorithm.

4 Methodology

Ontology merging is the process of building a new coherent ontology (called a merged ontology) from given input ontologies to provide a unified access on the domain. Principally, it requires to know which elements are equal to each other. This can be achieved by an ontology alignment method to detect the corresponding pairs. Many significant advances such as [1,8,21,22] have already been made for the automatic ontology alignment. Moreover, the most state-of-the-art matching systems participate in OAEI campaigns [1] have achieved promising results in several use cases. Therefore, we assume that the alignments are already determined by an existing tool in this research. Same as our assumption, using the pre-determined mappings has previously been applied in [13,16,18].

Below, a schematic of our holistic multiple ontologies merging framework is illustrated in Fig. 1. It gets as an input a set of ontologies with their alignments, and automatically generates a merged ontology with a set of output mapping between the merged and the input ontologies. In the preprocessing step, the input ontologies and the pre-determined mapping are imported into a repository. Afterwards, the elements of the n input ontologies will be clustered into the k clusters in the clustering step. Finally, the combination phase will be applied to combine the k created clusters into the merged ontology. Thus, instead of the quadratic merge process, this holistic approach needs k merge operations. Here the number of clusters is noticeably smaller than the number of input ontologies ($k \ll n$). Therefore, **to overcome the scalability problem** in this framework, rather than merging a large number of ontologies (n), we merge a small number of clusters (k). Once the merged ontology is created, the output

[1] http://oaei.ontologymatching.org/

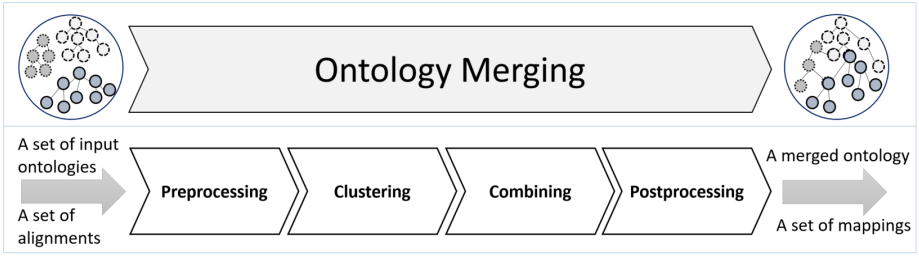

Fig. 1: Schematic of the holistic ontologies merging framework

mapping between the merge and the input ontologies will be produced through a backward process in the postprocessing phase.

Contributing to **RQ1**, the *clustering component* in Fig. 1 is accelerating the combination process by minimizing the search space in the way of breaking merging n ontologies into the merging of k clusters. Generally, clustering z elements of all input ontologies into k clusters needs $\frac{z(z-1)}{2}$ times comparisons. Alternatively, would be comparing $z-k$ elements with only k elements, which it requires $z \times k$ times comparisons. Therefore, we will use a *core-clustering* method, where all z elements will be compared with k core of the clusters through a semantic similarity function. We deem that the elements with a high number of correspondences in the input ontologies are more suitable to be considered as the cores. In this follow, the core of each cluster and the optimal number of clusters for each set of ontologies will be dynamically determined by using their mapping information.

To address **RQ2**, we divide the combination process into two steps through the *combining component* (Fig. 1): (i) *intra-combination*, and (ii) *inter-combination*. In the first step, the elements inside the clusters will be combined to create k sub-ontologies. In the *inter-combination* step, the k generated sub-ontologies will be attached to create the final merged ontology. Here, detecting the finest place of the join is a challenging task. We intend to find it in a heuristic method by narrowing comparison between the leaves and the core of sub-ontologies.

To capture the accuracy and consistency of the result, the predetermined merge requirements will be assessed before applying each combining process. The existing merge technique partially aims to satisfy some of them, however, they should be customized by the task at the hand (*RQ2-1*). To this end, we are providing a checklist including a variety of requirements (extracted from [13,15,16,18]), where the user can customize it (as is shown in Fig. 2). Each requirement will be performed as a set of rules with a weighting strategy (*RQ2-2*), in the way that they do not have a contraindicated with each other. Utilizing these requirements tends to guarantee the consistency of the created merged ontology, and consequence the quality of the merge result (*RQ2-3*). To accomplish consistency (*RQ2-4*), the conflicts and inconsistencies

should be detected at first via a reasoner, then they should be resolved. This is under our survey to handle inconsistencies with some conflict resolver such as subjective logic-based approach [12].

To develop a web-based tool, a preliminary version of $\mathcal{C}o\mathcal{M}erger$ is being provided by using the OWL API library [2] and a user-friendly GUI as is shown in Figs. 2 and 3. The further version might also include the analysis the execution log of the merge evaluator and visualization of the results.

Evaluation Protocol: We address the required evaluations as below:

E1. To study our main research question, we will compare the quality and complexity of merging multiple ontologies by a series of the binary merges rather than our holistic merge approach. The quality will be measured as states in *E5*, and the complexity will be measured as the number of required operations.

E2. The aim of clustering is to accelerate the merge process. Therefore, to evaluate *RQ1-1*, we will compare the quality of the merge result and complexity of the merge operations with and without using clustering.

E3. To evaluate our heuristic approach on *RQ1-2*, we will compare the quality of the merge result on the experimental tests for $k = 1, ..., n$. Besides, the cohesion and coupling [2] of the created clusters will be analyzed.

E4. A use case testing will investigate the feasibility of which requirements are worth being considered (*RQ2-1*) and to what extent they can be satisfied (*RQ2-2*).

E5. The *quality* of the merge result regarding *RQ2-3* can be evaluated in three scenarios: (i) Measuring the integrity of the merged ontology with the merge quality criteria, namely, compactness, completeness, and minimality [9]. In addition, we are investigating to recast these measures in a broad range of systematic criteria in our evaluator tool. (ii) Comparing our result with the mentioned state-of-the-art approaches. (iii) Comparison with human experts results on the part of our datasets.

E6. The *consistency* of the merge result can be evaluated by measuring the fulfillment of the customized merge requirements with either a reasoner or an expert (*RQ2-4*).

E7. We demonstrate the method's *scalability* by illustrating the performance test results on the set of real-life ontologies from BioPortal repository [3] and OAEI datasets. The first one currently contains more than 700 biomedical ontologies with thousand of classes, and the second one includes several domains such as biodiversity and ecology, anatomy and conference domains. Here, the runtime performance will be evaluated based on the number of ontologies versus the time required for merge operations.

[2] http://owlapi.sourceforge.net
[3] https://bioportal.bioontology.org

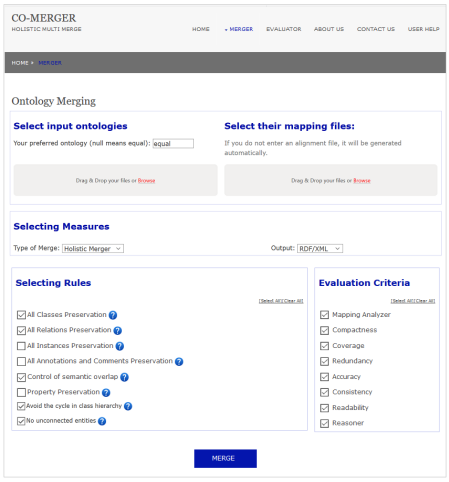

Fig. 2: Merger

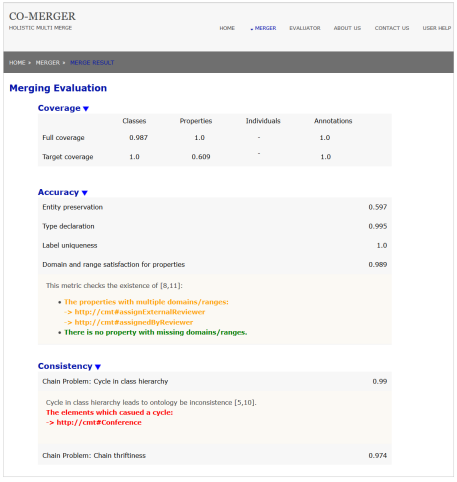

Fig. 3: Evaluator

5 Preliminary Results

The first conceptual ontology-based data integration workflow has already been represented in [3]. Additionally, we investigated the role of mapping in the merge process [5]. Moreover, we developed a highly accurate similarity method [4] by applying Information Content as an optimization problem [4]. We will revise it to be extended in our similarity function. Finally, the first version of our tools, *merger* (Fig. 2) and *evaluator* (Fig. 3) are under our development to be published online.

6 Discussion

Ontology merging is often a necessity in applications of the Semantic Web. To this end, our aim is to provide a holistic multiple ontologies merging, namely $\mathcal{C}o\mathcal{M}erger$, to satisfy the scalability issue. The efficient processing will be held by breaking the n ontologies processing into the k clusters merging, with a minor overhead of the clustering process. Each component has a high effect on the quality of the final merge result, therefore, the difficulty of this research would be carefully fine-tune the correctness of each sub-function. Besides, in the aspect of knowledge integration, the possibility of ambiguous knowledge being introduced in the final merged ontology will be another obstacle in this research, which bring us to deal with the reasoning from ambiguous knowledge challenge [19]. The future works of this research can be extended to the merging data in the schema-level on the Linked Open Data (LOD) scenarios, also utilizing parallel techniques in our framework.

[4] http://simbio.uni-jena.de

Acknowledgments

The author would like to thank Prof. Birgitta König-Ries and Dr. Alsayed Algergawy for their valuable supervising. The author is supported by a scholarship from German Academic Exchange Service (DAAD).

References

1. A. Algergawy, S. Babalou, M. J. Kargar, and S. H. Davarpanah. Seecont: A new seeding-based clustering approach for ontology matching. In *ADBIS*, 2015.
2. A. Algergawy, S. Babalou, and B. König-Ries. A new metric to evaluate ontology modularization. In *SumPre with ESWC*, 2016.
3. S. Babalou, A. Algergawy, and B. König-Ries. An ontology-based scientific data integration workflow. In *Proc. of the 29th GVDB.*, pages 30–35, 2017.
4. S. Babalou, A. Algergawy, and B. König-Ries. A particle swarm-based approach for semantic similarity computation. In *OTM*, pages 161–179. Springer, 2017.
5. S. Babalou, A. Algergawy, B. Lantow, and B. König-Ries. Why the mapping process in ontology integration deserves attention. In *Proc. of the 19th iiWAS Conf.*, pages 451–456. ACM, 2017.
6. C. Batini, M. Lenzerini, and S. B. Navathe. A comparative analysis of methodologies for database schema integration. *In CSUR*, 18(4):323–364, 1986.
7. O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.
8. S. H. Davarpanah, A. Algergawy, and S. Babalou. Fuzzy inference-based ontology matching using upper ontology. In *ADBIS*, 2015.
9. F. Duchateau and Z. Bellahsene. Measuring the quality of an integrated schema. In *ER*, pages 261–273, 2010.
10. M. Fahad. Merging of axiomatic definitions of concepts in the complex owl ontologies. *AIR*, 47(2):181–215, 2017.
11. W. Hu, J. Chen, H. Zhang, and Y. Qu. How matchable are four thousand ontologies on the semantic web. *In ESWC*, 2011.
12. A. Jøsang and S. J. Knapskog. A metric for trusted systems. In *Proc. of the 21st National Security Conf. NSA*, 1998.
13. S. P. Ju, H. E. Esquivel, A. M. Rebollar, M. C. Su, et al. Creado–a methodology to create domain ontologies using parameter-based ontology merging techniques. In *MICAI*, pages 23–28. IEEE, 2011.
14. M. Mahfoudh, L. Thiry, G. Forestier, and M. Hassenforder. Algebraic graph transformations for merging ontologies. In *MEDI*, pages 154–168. Springer, 2014.
15. N. F. Noy and M. A. Musen. The prompt suite: interactive tools for ontology merging and mapping. *IJHCS*, 59(6):983–1024, 2003.
16. R. A. Pottinger and P. A. Bernstein. Merging models based on given correspondences. In *Proc. of the 29th int. conf. on VeLDB-Volume 29*, pages 862–873, 2003.
17. E. Rahm. The case for holistic data integration. In *ADBIS*, pages 11–27, 2016.
18. S. Raunich and E. Rahm. Target-driven merging of taxonomies with atom. *Inf. Syst.*, 42:1–14, 2014.
19. S. K. Reed and A. Pease. Reasoning from imperfect knowledge. *Cognitive Systems Research*, 41:56–72, 2017.
20. K. Saleem, Z. Bellahsene, and E. Hunt. Porsche: Performance oriented schema mediation. *Information Systems*, 33(7):637–657, 2008.
21. M. Shamsfard, B. Helli, and S. Babalou. Omega: Ontology matching enhanced by genetic algorithm. In *ICWR*, 2016.
22. P. Shvaiko and J. Euzenat. Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176, 2013.
23. G. Stumme and A. Maedche. Fca-merge: Bottom-up merging of ontologies. In *IJCAI*, pages 225–230, 2001.