How good is this merged ontology?

Samira Babalou* [1], Elena Grygorova [1], Birgitta König-Ries [1 2]

[1]Heinz-Nixdorf Chair for Distributed Information Systems
Institute for Computer Science, Friedrich Schiller University Jena, Germany
[2]Michael-Stifel-Center for Data-Driven and Simulation Science, Jena, Germany
{samira.babalou,elena.grygorova,birgitta.koenig-ries}@uni-jena.de

Abstract. With the growing popularity of semantics-aware integration solutions, various ontology merging approaches have been proposed. Determining the success of these developments heavily depends on suitable evaluation criteria. However, no comprehensive set of evaluation criteria on the merged ontology exists so far. We develop criteria to evaluate the merged ontology. These criteria cover structure, function and usability of the merged ontology by evaluating General Merge Requirements (GMR)s, analyzing the intended use and semantics, and considering the ontology and entity annotation, respectively. We demonstrate the applicability of our criteria by providing empirical tests.

Keywords: Semantic Web . Ontology evaluation . Ontology merging

1 Introduction

Merging ontologies involves identifying correspondences among the entities in the different ontologies and combining them into a new merged ontology. Given the central role these merged ontologies play in realising real world applications, such as knowledge reusing [1] and query processing [2], there is a strong need to establish evaluation methods that can measure their quality. Existing studies on evaluation of the merged ontology suffer from various drawbacks as detailed in Sec. 2. Automatic merge evaluation can support an expert evaluation along a broad and customizable range of criteria in different aspects.

We adapt evaluation dimensions from well-known ontology evaluation frameworks [3, 4] in the context of ontology merging, formulate our evaluation criteria on top of the categories proposed there classified into structural, functional, and usability-profiling measures, and analyze how these dimensions can be evaluated on the merged ontology in practice. Our final contribution is an online ontology merging evaluator, the $\mathcal{CoMerger}$ tool, which is independent of any merge method.

2 Literature Review

Most ontology merging approaches lack sufficient experimental evaluation on the merged result (cf. [5]). Other ontology merging studies, such as GCBOM [6]

evaluate in terms of the size of created merged ontologies, only. The state of the art is far from providing an adequate benchmark. While [7] provides a benchmark, it includes simple taxonomies only and just compares the number of paths and concepts of a tool-generated merged ontology to a human created one. The benchmark proposed in [8] includes only few and small ontologies and focuses on criteria tailored to the GROM tool [9]. Moreover, user-based evaluation is a complex, time-consuming, and error-prone task. This concludes a need for a comprehensive evaluation, to which we contribute.

3 Proposed Quality Criteria for Evaluating of the Merged Ontology

An ontology merging evaluator measures the quality of the merged ontology $\mathcal{O}_M$ based on a set of source ontologies $\mathcal{O}_S$ and their mapping $\mathcal{M}$ with respect to a set of evaluation criteria. To evaluate the merged ontology in a systematic way, we adapt successful evaluation dimensions from two ontology evaluation frameworks [3, 4] and customize them in the context of ontology merging. These two works introduced structural and functional evaluation dimensions. Moreover, in [4] the usability-profiling and in [3] the reliability, operability, and maintainability dimensions are presented. Since the last three mentioned dimensions are not affected by the merge process, we mainly focus on structural, functional and usability-profiling dimensions. We build our criteria on top of these classifications, as follow:

(1) Measuring the structural dimension. It focuses on syntax and formal semantics. In this form, the topological and logical properties of an ontology may be measured by means of a metric. To classify the criteria in this dimension, we use the classification of [10], which distinguishes into three dimensions (integrity, logic properties, and model properties) to structure our list of twenty General Merge Requirements (GMR)s.

This list has been build by reviewing publications in three different research areas, including ontology merging methods, benchmarks, and ontology engineering and extracting relevant criteria. Thus, it comprehensively considers all topological properties. Moreover, we consider the consistency aspect from [11], as suggested in [3].

(2) Measuring the functional dimension. This dimension is related to the intended use and semantics of a given merged ontology and of its components. Functional measures have been quantified by precision $P = \frac{|TP|}{|TP|+|FP|}$ and recall $R = \frac{|TP|}{|TP|+|FN|}$ in [4], where, TP=True Positive, FP=False Positive, and FN=False Negative. This definition is adapted by choosing an appropriate domain for positive and negative responses from the matching between the ontology structure and the intended usage and meaning. High (low) precision and recall label $\mathcal{O}_M$ as GOOD (WORSE). Low precision and high recall make it LESS GOOD, and vice a versa BAD.

An intended conceptualization corresponds to the expertise of an ontology's intended users [4]. The expertise boundary is provided by the task that should

be accomplished with the help of the ontology or at least the schema of that expertise that should be captured. Since expertise is by default in the cognitive "black-box", ontology engineers have to elicit it. Thus, precision and recall of an ontology graph can be measured against experts' judgment, or a data set assumed as a qualified expression of experts' judgment. We find two scenarios to accomplish it:

(i) Using Competency Questions. One of the approaches in [4] to capture the intended use of an ontology is to use Competency Questions (CQ)s. A set of CQs is complete in the sense that if the ontology can provide correct answers to all questions, then the ontology can serve its intended purpose. We determine them in the context of the merged ontology w.r.t. the source ontologies. Thus we define TP, FP, and FN based on the expected answers of the source ontologies.

(ii) Using query scenario. Comparing the individuals and is-a relations queries from merged $\mathcal{O}_M$ and source $\mathcal{O}_S$ ontologies can provide the environment to capture the intended semantic. Thus, we provide a list of queries which the $\mathcal{O}_S$ can or cannot answer, and then compare with the achieved answers from $\mathcal{O}_M$.

(3) Measuring the usability-profile. It focuses on the ontology profile to address the communication context of an ontology. We measure:

- *Annotation about the ontology itself*: It evaluates the existence and correctness of (1) ontology URI, (2) ontology namespace, (3) ontology declaration, and (4) ontology license (requiring modeling compatibility of different licences).
- *Annotation about ontology's entities*: This includes: (1) Label uniqueness to observe whether the created labels are unique [12]. (2) Unify naming to evaluate whether all entity's names follow the same naming conventions in the merged ontology [13]. (3) Entity type declaration to check whether these entities have been explicitly declared [13].

4 Empirical Analysis: Assessments in practice

The introduced criteria have been implemented in our merge framework *CoMerger* [14] and distributed under an open-source license along with publishing the used merged ontologies. The used patterns and exact algorithms to detect and repair each GMR have been documented in our portal[1]. For the consistency test, we refer to [11]. We have selected[2] a set of well-known ontologies with their available mapping, and created merged ontologies, by combining the corresponding entities and reconstructing their relations, and evaluate our criteria on them. The result of evaluating the structural and usability-profile dimensions are available in our repository[3]. In this paper, we demonstrate the evaluation of the functional dimension as applicability of our method:

[1] http://comerger.uni-jena.de/requirement.jsp
[2] Datasets: https://github.com/fusion-jena/CoMerger/tree/master/EvaluationDataset
[3] https://github.com/fusion-jena/CoMerger/blob/master/EvaluationDataset/result.md

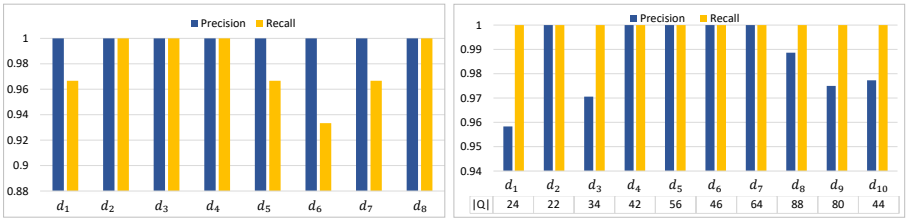

Fig. 1: Left: Functional measure's evaluation for the intended use via CQs; Right: Functional measure's evaluation for intended semantics via queries. Considering high values above 0.5, all tested ontologies achieve "GOOD" labels.

(a) Test with CQs. By analyzing user-provided Competency Questions, we aim to observe how the ontology's structure is aligned with the intended use of the created merged ontology. We have provided a list of CQs (available in our portal) for the conference domain datasets. To quantify the precision and recall, we determine positive and negative CQs along with TP, FN, and FP. The *positive CQ* is a CQ that at least one of $\mathcal{O}_S$ can answer. For a *negative CQ*, none of the $\mathcal{O}_S$ can answer it. If $\mathcal{O}_M$ correctly answers a positive query, we mark it as TP, and if it incorrectly answers it, we mark it with FN. If $\mathcal{O}_M$ provides a correct (wrong) answer to a negative query, we mark it as FP (TN). The results are demonstrated in Fig. 1, left, where precision and recall are shown for each dataset. All $\mathcal{O}_M$ evaluated in this test achieved precision 1 because the FP of all of them is zero. If none of the $\mathcal{O}_S$ can answer the negative CQs, the $\mathcal{O}_M$ in our test could not answer it, since no further information than $\mathcal{O}_S$ is added to the $\mathcal{O}_M$ during the merge process. If an $\mathcal{O}_M$ is built by human intervention, that might bring some new knowledge. In this case, non-zero values would be possible for FP. As a whole, the recall of all tested ontologies varied between 0.93 and 1.

(b) Test with queries. To evaluate the intended semantics of the merged ontology, we created two types of queries on individuals and is-a relations queries. In the is-a-based queries, for each subclass-of relation like *'*$A \sqsubseteq B$*'*, we make a true query *'*$A \sqsubseteq B$*?'*, and a false query like *'*$A \sqsubseteq C$ *?'*. For each individual c of concept A, we create a positive and negative individual query like *'is c a A?'* and *'is c a B?'*. In both, *'*$B \neq C$*'* and *'*$A \not\sqsubseteq C$*'*. We expect that the answer from $\mathcal{O}_M$ for the true query is true and for the false query is false. If so, we mark them as intended answers. Otherwise, we mark it as non-intended answers. If $\mathcal{O}_M$ correctly (wrong) answers a non-intended answer, we mark it as FP (TN). If $\mathcal{O}_M$ correctly (wrong) answers an intended answer, we mark it as TP (FN). Fig. 1, right shows the precision and recall of results from running 500 queries on our used datasets. The test demonstrates that the intended semantics is high.

5 Conclusion

This paper contributes to providing the multi-aspects of evaluating the quality of a merged ontology w.r.t. its source ontologies into structural, functional and usability-profiling dimensions. A practical assessment has been presented. The use case scenario evaluation and meta-evaluation are on our future agenda.

Acknowledgments

S. Babalou is supported by a scholarship from German Academic Exchange Service (DAAD).

References

1. M. T. Finke, R. W. Filice, and C. E. Kahn Jr, "Integrating ontologies of human diseases, phenotypes, and radiological diagnosis," *J AM MED INFORM ASSN*, vol. 26, no. 2, pp. 149–154, 2019.
2. K. M. Livingston, M. Bada, W. A. Baumgartner, and L. E. Hunter, "KaBOB: ontology-based semantic integration of biomedical databases," *BMC bioinformatics*, 2015.
3. A. Duque-Ramos, J. T. Fernández-Breis, R. Stevens, N. Aussenac-Gilles, *et al.*, "OQuaRE: A SQuaRE-based approach for evaluating the quality of ontologies," *JRPIT*, vol. 43, no. 2, p. 159, 2011.
4. A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann, "Ontology evaluation and validation: an integrated formal model for the quality diagnostic task," *On-line: http://www. loa-cnr. it/Files/OntoEval4OntoDev_Final. pdf*, 2005.
5. 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*, pp. 23–28, IEEE, 2011.
6. M. Priya and C. A. Kumar, "An approach to merge domain ontologies using granular computing," *Granular Computing*, pp. 1–26, 2019.
7. S. Raunich and E. Rahm, "Towards a benchmark for ontology merging.," in *OTM Workshops*, vol. 7567, pp. 124–133, 2012.
8. M. Mahfoudh, G. Forestier, and M. Hassenforder, "A benchmark for ontologies merging assessment," in *KSEM*, pp. 555–566, 2016.
9. M. Mahfoudh, L. Thiry, G. Forestier, and M. Hassenforder, "Algebraic graph transformations for merging ontologies," in *Int. Conf. on Model and Data Eng.*, pp. 154–168, Springer, 2014.
10. S. Babalou and B. König-Ries, "GMRs: Reconciliation of generic merge requirements in ontology integration," *In SEMANTICS Poster and Demo.*, 2019.
11. S. Babalou and B. König-Ries, "A subjective logic based approach to handling inconsistencies in ontology merging," in *OTM*, Springer, 2019.
12. N. F. Noy and M. A. Musen, "The PROMPT suite: interactive tools for ontology merging and mapping," *IJHCS*, vol. 59, no. 6, pp. 983–1024, 2003.
13. M. Poveda-Villalón, A. Gómez-Pérez, and M. C. Suárez-Figueroa, "Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation," *IJSWIS*, vol. 10, no. 2, pp. 7–34, 2014.
14. S. Babalou, E. Grygorova, and B. König-Ries, "CoMerger: A customizable online tool for building a consistent quality-assured merged ontology," in *In 17th Extended Semantic Web Conference (ESWC'20), Poster and Demo Track*, June, 2020.