Measuring the Effect of a Guideline-based Training on Ontology Design with a Competency Questions based Evaluation Approach

Martin Boeker [1]*, Niels Grewe [2], Johannes Röhl [2], Daniel Schober [1],
Stefan Schulz [1,3], Djamila Seddig-Raufie [1], and Ludger Jansen [2]

[1] Institute of Medical Biometry and Medical Informatics,
University Medical Center Freiburg, Germany
[2] Institute of Philosophy, University of Rostock, Germany
[3] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria

Abstract:

OBJECTIVE: (a) To measure the effect of a guideline-based training on the performance of ontology developers compared with the performance after unspecific training by a competency question based evaluation; and (b) to provide empirical evidence for the applicability of competency questions in formal ontology evaluation in general.

BACKGROUND: A close connection between ontology development and ontology evaluation as quality management procedure can been attained with the use of competency questions. Competency questions are often used as a semi-formal specification of requirements for an ontology under development. Hence they can also be used as evaluation instruments, in order to check how far an ontology fulfills these requirements.

METHODS: A randomized controlled trial was conducted with two groups of 12 students each. The intervention consisted in a differential guideline-based training on ontology development vs. unspecific training. After a group-specific training focusing on three topics per group, performance of students was assessed with 12 exercises (2 exercises per topic), in which the students had to apply their skills. Different types of competency questions were elaborated for the analysis of the students' ontologies. We used the proportion of correct answers as a measure of ontology quality.

RESULTS: On single topic level, the performance of ontology developers increased after guideline-based training for two out of the six topics: it increased from a proportion of 0.46 to 0.63 for the topic *Process & Participation* and from 0.44 to 0.53 for the topic *Collective Material Entity*. In regression analysis, a positive correlation was shown between the performance of students on untrained topics and the performance after specific guideline-based training. Moreover, in multiple regression analysis an overall effect of specific training of 0.09 was calculated ($p < 0.1$).

CONCLUSION: The results show an effect of a specific guideline-based training on the performance of ontology developers compared to the performance after unspecific training by an increase of about 10 % on the rate of correct competency questions. In addition, this study has shown the general applicability of competency questions in a formal ontology evaluation scenario. However, the study also shows that the training of ontology developers and their performance evaluation is a tedious task. The resulting performance of ontology developers is more dependent on the *a priori* individual competencies than on the specific acquired skills after training.

*To whom correspondence should be addressed: martin.boeker@uniklinik-freiburg.de

1 Introduction

Evaluation of ontology quality is important for scientific and engineering communities who develop or apply them. Various methods have been suggested as measures or indicators for ontology quality [HDG12]. However, many of them are appropriate for specific use cases only and cannot easily be transferred to general ontology evaluation [Euz07, EMS$^+$11, VS07]. Other methods have been suggested both to evaluate and improve the quality of ontologies [GW02], but they cannot be automatized or do not yield outcome parameters that can easily be interpreted [NVB$^+$13].

In many cases, ontology evaluation is requested because of the need to objectify whether an ontology has to be improved. In this case, the evaluation results should also give a hint on where improvement is most needed and most effective. In our view, the closest connection between ontology development and ontology evaluation as quality management procedure can been achieved by using competency questions. Competency questions are often formulated as a semi-formal specification of requirements for an ontology under development [GF95]. Hence they can also be used as evaluation instruments, in order to ascertain how far an ontology fulfills these requirements. In this way competency questions can be seen as a bracket that encloses the complete life cycle of an ontology, indicating to what extent the *intended* functions and requirements are met by the actual development stage.

To a certain extent, the ontology development community can learn from other engineering disciplines, esp. computer science, to apply the large corpus of methodology developed in these domains for the benefit of ontology development and evaluation. One approach used in software programming is test-driven development, which relies on so called unit tests [VG06]. With competency questions, ontology development frameworks become feasible, which integrate the testing for their fulfillment at every step of the development cycle and, thus, make quality assurance a standard step in development process.

Twenty years ago, Fox and Gruninger introduced competency questions to specify requirements and evaluate resulting ontologies [FCF93, GF94]. To our knowledge, no empirical data from randomized trials is available, which could provide evidence for the applicability of competency questions in real-world scenarios.

The aim of this paper is (a) to measure the effect of guideline-based ontology training on the performance of ontology developers compared with the performance after unspecific training by a competency question based evaluation; and (b) to provide empirical evidence for the applicability of competency questions in formal ontology evaluation in general.

1.1 Competency question for the specification and evaluation of ontology

With the aim to formally represent the activities and structures of business enterprises, ontologies representing general things like time, causality, activity, and constraints were defined in the TOronto Virtual Enterprise project (TOVE). The need for an evaluation of the representation was addressed by Fox [FCF93], who introduced a proposal for an evaluation framework of formal representations based on requirements and *competencies*

group	domain	topic
1	upper-level ontology upper-level ontology ODP	Process & Participation Immaterial Object Closure ODP
2	upper-level ontology upper-level ontology ODP	Collective Mat. Entity Information Object Spatial Disjointness ODP

Table 1: Topics used for interventional training of the two student groups. The topics correspond to modules in the guideline-based curriculum. Each group received specific training on the indicated topics and unspecific training on the remaining three topics. Strictly speaking, this educational design uses *two* different interventions due to context dependency of the different topics. Both interventions apply the same instructional method (guideline-based training). For each topic the students were asked to develop two ontologies in the assessment phase of the experiment.

on different levels of the representation. For the first time, these authors introduced the concept of *competency questions* in the domain of ontology development. The term competency (assessment) question had already been in use for outcome assessment in educational sciences and as a legal term of art. It was eventually suggested as an instrument for the evaluation of software and software systems by Gruninger [GF94].

Competency questions have been used in a systematic way during the development of the TOVE Traceability Ontology [KFG99]. This ontology was constructed using competency questions at different steps and levels of the development process. TOVE's logical axioms were checked by a theorem prover against given competency questions.

Gangemi et al. investigated competency questions for ontology evaluation at a theoretical level and included competency questions as one of the crucial parts into a ”comprehensive model for ontology evaluation and validation” [GCC$^+$05, GCCL06].

Obrst stresses the use of competency questions to specify requirements towards an ontology in the requirements analysis [OCM$^+$07]. As a result, the same competency questions can then be used in the design and evaluation phase to test if the ontology fulfills these requirements [SSSS01]. Although the importance and feasibility of competency questions in ontology evaluation has been stressed, there is no evidence of more recent empirical research on real-world ontology evaluation with competency questions.

2 Methods

2.1 Design and analysis of the study

This is a study on ontology training and evaluation which was conducted in September 2011 with 24 participants. The curricular and experimental setting has been described in detail elsewhere [BSR$^+$12, BJG$^+$13]. Briefly, the study was planned as an educational trial, in which a guideline-based[1] curriculum on biomedical ontology development was taught differentially to students, so that the effect of the training on the students performance could be measured.

The *study design* was a randomized controlled (parallel) group design with a complex crossed intervention (see below). We invited students with a subject in the life sciences and proven knowledge in computer science to participate in the study, by direct contact with the teachers of their faculty. 24 students from Austria, Germany, Slovenia, and Switzerland participated in the study. They were randomly balanced allocated to two groups for differential training on ontology design after three days of shared basic training.

The background for the *intervention* was a curriculum for ontology development based on a good practice guideline[1] [BSR$^+$12], addressing important aspects of ontology theory and ontology engineering practice: foundations of philosophical and formal ontology, syntax and semantics of the description logics OWL DL, application of top-level categories, and ontology design patterns (ODPs). The curriculum was modularized and designed for learning with exercises.

The study started with three full days of general training sessions delivered to all participants. After this followed the intervention proper that consisted of training on specific topics delivered to one group only, compared with training unspecific to these topics for the other group. Thus, for several modules students were either instructed to solve a certain modelling problem according to the guideline, or received a training that was unspecific with regard to the solution of this problem (Table 1). Topics selected for the interventional part of the study were derived from the guidelines' parts on upper level ontologies [MBG$^+$03, BSSH08, Se12] and ontology design patterns [SSM$^+$11].

The *primary outcome* of the study was the performance of students in ontology development measured as the distance of the ontology artefacts they produced compared with gold standard ontologies produced by experts, using ontology similarity metrics. The methods and results of these outcome measures have been presented elsewhere [BJG$^+$13].

A *secondary outcome* measure was defined as as the proportion of correct answers to pre-formulated competency questions. The present paper focuses on this secondary outcome measure. For each of the six interventional modules, two exercises had been prepared by the author team. Each student's task was to develop twelve test ontologies based on these exercises. Six of these ontologies concerned topics on which they had received specific guideline-based training, and six ontologies concerned topics for which they only received unspecific training. For each ontology development exercise, students were provided with a short written text defining the requirements and, as a starting point, an OWL file with a list of domain classes to be elaborated on, together with a fragment of the upper-domain ontology BioTopLite. The fragment of BioTopLite provided top-level categories and relations, enriched by constraining axioms.

Together with the gold-standard ontologies, we had created a set of suitable competency questions for each exercise in a consensus process. *Data collection* was done by comparing result sets for competency questions from the Protégé DL query tab with the given result sets in the documentation of the competency questions. Answers were coded as 0 for a

[1] http://purl.org/goodod/guideline

wrong answer and 1 for a correct answer.

All *statistical analyses* were performed with STATA 12.1.[2] As the first step of analysis, an item analysis was performed on topic level to estimate the reliability of the competency question set for this topic and to reduce the set to competency questions with a positive item correlation. All further results were calculated on the basis of the reduced item sets (see Table 2). For each participant the proportion of correct answers was calculated. T-tests were performed to compare the mean of the trained and untrained students groups on topic level. A Bonferroni-adjusted significance level was calculated to correct for multiple testing on individual topic level: $\alpha = 0.0167$. Differences of means and Cohens d effect sizes were estimated as effect measures.

After item analysis on group level with three topics aggregated respectively, a t-test was performed to analyse the training effect for each group regardless of the topic. Regression analysis was used to estimate the correlation between results with unspecific training and after group-specific training for both interventional groups. To computationally isolate the effect of the specific guideline-based training vs. unspecific training from other independent variables and confounders *for both groups* a multivariate mixed random effects model was calculated. The type of intervention and the topic of the ontology were used as adjustment parameters in this model.

Participants received an expense allowance of 500 € for their participation in the training and the study. Before their agreement, students had been informed about all details of the curriculum and the following study. As part of the agreement, it was explicitly stated that the payment of the allowance was dependent on the students' complete attendance and full cooperation during the training sessions and the study but not on their success in the assessments or answers in the questionnaires.

Ethical approval was requested from the ethical authority of the University of Freiburg, Freiburg, Germany. The chair of the University of Freiburg ethics committee reviewed the project and concluded that a full formal ethics committee statement was not required due to the educational nature of the study. It was designed according to the general requirements for educational studies at the Freiburg University Medical Center, and was performed with informed written consent of the participants.

2.2 Types of competency questions

Competency questions were formulated in Manchester OWL 2 Syntax as DL axioms. Gold standard ontologies provided by experts fulfilled all competency questions. In this way competency questions and exercises were adapted; only such classes and relations were included in the exercise stub ontologies which were included in the competency questions and gold standard ontologies respectively. For each of the competency questions it was documented on how to assess the complete competency question with the DL query tab of Protégé: (a) check for subclasses (SubclassOf) with or without descendant classes, (b) check for equivalent classes (EquivalentTo), or (c) check for superclasses with or without

[2] StataCorp LP, http://www.stata.com

ancestor classes. The expected result set of the competency question against the target ontology was provided as a subset of classes from the test exercise class set. This result set could also be empty or return the value *Nothing* in case the competency question logically contradicted the assertions of the target ontology.

The competency questions used in this study can be classified into different categories which will be introduced here by way of example.

On the lowest level of ontological complexity, the correctness of the asserted hierarchy was checked by a substantial number of competency questions, e.g. for the correct hierarchical insertion of the class *BiologicalMembrane* in an exercise for the Spatial Disjointness ODP.

> *BiologicalMembrane* SubclassOf *CellOrganelle*

The correctness of the asserted hierarchy was tested with a series of competency questions of this type. We frequently relaxed the restriction for correctness on direct subclasses when only the correct order of classes was important.

The correct and exhaustive introduction of disjointness axioms was tested as an unsatisfiable conjunction of the disjoint classes:

> (*CellOrganelle* and *Cytoplasm*) EquivalentTo *owl:Nothing*

The correct usage of classes and relations of the top-level ontologies was the largest part of the exercises. In these exercises, the correct top-level classes and relations had to be chosen to fulfill the competency questions. Therefore, resulting ontologies were checked against a series of competency questions that axiomatized these requirements. For example, intervention group 1 was trained for representation of processes. The corresponding exercise was checked against the following competency questions.

> *Diagnosing* SubclassOf *Action*
>
> *Diagnosis* SubclassOf *InformationObject*
>
> *Diagnosing* SubclassOf (**hasAgent** some *Physician*)
>
> *Diagnosing* SubclassOf (**hasPatient** some *Patient*)
>
> *Diagnosing* SubclassOf (**hasOutcome** some *Diagnosis*)
>
> *Radiologist* SubclassOf (**agentIn** some *ImagingDiagnosing*)

The correct usage of ODPs was tested against axioms prescribed for this design pattern, e.g. in the case of a spatial disjointness pattern as it had been taught to group 2 (see Table 2).

> *Mitochondrium* and (**hasLocus** some *GolgiApparatus*)
> EquivalentTo *owl:Nothing*

For the Closure ODP as trained with group 1 the complete pattern was checked using the following two competency questions:

ChildDenture and not (**hasPhysicalPart** some *IncisorTooth*)
EquivalentTo *owl:Nothing*

ChildDenture and (**hasPhysicalPart** some *CanineTooth*)
EquivalentTo *owl:Nothing*

By way of using DL queries, we could assure that syntactically different but logically equivalent modelling solutions retrieved the same result sets in the evaluation.

3 Results

In this study on the effect of a guideline-based training on the performance (skills) of students in ontology development we used competency questions to measure the performance of the students. Two groups of students (n=24) received guideline-based training on three topics each and received only unspecific training on the other topics respectively. Each student (n=24) was tested on 6 topics with two exercises in which they had to develop an ontology (12 test ontologies from each student).

We formulated between 17 and 29 competency questions for each of the topics (see Table 2), i.e. aggregated for two ontologies. On the full set of competency questions (items) prior to item deletion Cronbachs alpha indicated a wide range of reliability of the tests for each topic (0.81 to 0.03). After deletion of items with negative item reliability, the competency question sets were reduced to sizes between 13 and 25 items for each individual topic measure. Reliability was improved by the deletion of items, esp. for the topics with low initial reliability from 0.03 to 0.5 for *Immaterial Object* and from 0.69 to 0.77 for *Collective Material Entity* (Table 2). In this study with only a small number of participants, the individual tests for each topic were characterized by good reliability.

The measures presented in the following reflect the proportion of correctly answered competency questions of all competency questions. The mean of the proportion of correct competency questions for the *untrained* group was between 0.44 (*Collective Material Entity*) and 0.58 (Spatial Disjointness ODP). The mean for the trained group was between 0.51 (*Information Object*) and 0.63 (*Process & Participation*). The standard deviation (SD) ranged from 0.12 to 0.23 for the untrained group and from 0.14 to 0.26 for the trained group. The higher SD in the trained group is a measure for the higher variance in this group which is an indicator of a training effect.

Of much more interest for the estimation of a training effect is the difference between trained and untrained groups. Only for two out of six topics a relevant difference of means between trained and untrained groups were found: 0.17 with a pooled SD of 0.25 for Process & Participation and 0.09 with an SD of 0.18 for Collective Material Entity. For the Spatial Disjointness ODP the difference was 0.03 and for Immaterial Object it was even -0.01. Standardized to SD, these differences result in Cohen's d effect sizes of 0.77 for *Process & Participation* and 0.54 for *Collective Material Entity*. Differences between trained and untrained groups are not significant on a Bonferroni-adjusted significance level of $\alpha = 0.0167$. Summarized for this part of the analysis, the measurement showed a

Parameter	group 1 trained Process & Participatn.	 Immaterial Object	 Closure ODP
Number CQ	17	23	19
Cronbach's α CQ	0.75	0.03	0.83
Number CQ reduced	15	15	19
Cronbach's α reduced	0.79	0.50	0.83
Mean untrained group	0.46	0.56	0.55
SD untrained group	0.20	0.15	0.23
Mean trained group	0.63	0.55	0.55
SD trained group	0.26	0.14	0.23
Mean difference	0.17	-0.01	0.0
SD Pooled	0.25	0.14	0.22
Cohen's d effect size	0.77	-0.08	0.0

Parameter	group 2 trained Collective Mat. Entitiy	 Information Object	 Spatial DJ ODP
Number CQ	29	27	17
Cronbach's α CQ	0.69	0.81	0.62
Number CQ reduced	23	25	13
Cronbach's α reduced	0.77	0.81	0.63
Mean untrained group	0.44	0.50	0.58
SD untrained group	0.12	0.14	0.15
Mean trained group	0.53	0.51	0.61
SD trained group	0.21	0.22	0.21
Mean difference	0.09	0.0	0.03
SD Pooled	0.18	0.18	0.18
Cohen's d effect size	0.54	0.02	0.15

Table 2: Results for competency questions (CQs) grouped by topic. Number of CQs and corresponding Cronbachs α are presented for each topic, before and after items with negative item correlation were deleted. Further results are based on the reduced item set. In the upper part of the table, the results are presented for the topics for which group 1 was specifically trained. In the lower part, the results for which group 2 received specific training. Only the results for *Process & Participation* and *Collective Material Entity* indicate relevant differences and effect sizes between untrained and trained groups. Differences are not significant on a Bonferroni-adjusted significance level of $\alpha = 0.0167$. SD = standard deviation; ODP = ontology design pattern.

relevant difference between specifically trained and untrained students for two out of six topics, however, these differences were not statistically significant.

If analyzed on group level by aggregating three topics for each group, the difference between the two groups was 0.10 for those topics on which group 1 had received specific training, and -0.01 for those topics on which group 2 had been trained. Neither of these differences are significant. Differences on topic level (Table 2) are blurred by the aggrega-

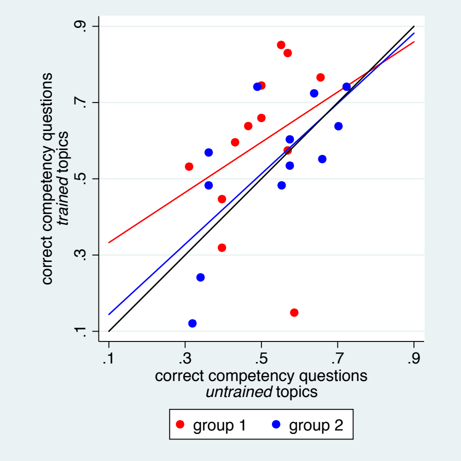

Figure 1: Average results for the proportion of correct competency questions of trained vs. untrained topics, color-coded for both groups. Be aware that the topics for the axes are different between groups in this figure. Most of the points are located above the symmetry line, indicating an overall training effect. The correlation between the students' performance without specific training and the performance after specific training is significant for group 2.

tion on group level and are no longer significant due to the limited power of the study. The correlation between the skills of students for untrained topics and specifically trained topics is depicted in Figure 1. There is a significant correlation between results for untrained topics and trained topics for group 2 (p=0.01; adjusted R^2=0.45) but no significant correlation for group 1. Student performance after specific training is (mainly) dependent on the value of their performance without specific training: students with weak performance on the untrained topics have low rates of correct competency questions after training and vice versa.

How large was the effect of the training on the performance of students in this experiment? Integrating all results in a mixed model regression analysis, a significant effect of 0.09 between the untrained and trained groups could be isolated (p=0.07). Figure 2 shows this effect of the training on both groups by plotting the normalized results of the trained vs. the untrained topics into one plot.

4 Discussion

In this study, we have shown a positive effect of guideline-based training on the performance of ontology developers compared with the performance after unspecific training by a competency questions based evaluation. Thus, we could provide evidence for the applicability of competency questions in formal ontology evaluation on the level of a randomized controlled trial.

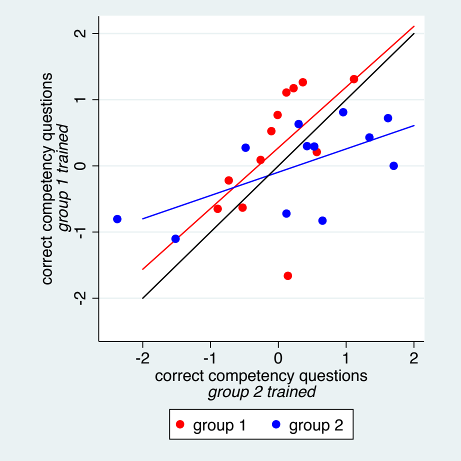

Figure 2: Results for the competency questions of trained vs. untrained topics for both groups. Depicted are Z-transformed results based on the means and SD of the untrained groups, respectively. The figure shows that students perform on average slightly better for topics they were trained for, as the students in group 1 tend to be above the diagonal, i.e. tend to score higher on the questions group 1 was trained for, whereas the students in group 2 tend to be below the diagonal, i.e. tend to score higher on the questions group 2 was trained for. However, the inter-individual variance dependent on the individual students performance is much higher than the training effect. The three students with extreme weak results nearest to the origin of the coordinate system have a strong influence on the group comparisons because of the small sample size.

The presented study has a number of limitations.

- A relevant effect of the guideline-based training could be shown only for two out of six topics in the group comparison on the *topic* level (Table 2). Effects were not statistically significant after Bonferroni correction for multiple testing. We consider a number of reasons responsible for this observation. On the one hand, we wanted to measure the effect of a complex cognitive intervention with many confounding influences which may result in considerable variance. On the other hand, we were bound to a small number of participants by our resources.

- The variability of the effects between topics can be explained by the very limited time for learning and training in contrast to the rather complex and difficult learning matter. For abstract topics like *Immaterial Object* or *Information Object* the transfer of newly acquired knowledge and skills into the practice of ontology development had been particularly difficult. To improve the design of our curriculum, more time must be allocated for training and consolidation of new knowledge and skills.

- Possibly, more precise insights into the training effect could have been achieved by a pure parallel group study design. However, our more complex design is superior to a parallel group design in two aspects: First, for practical and ethical reasons we could not have left a complete group without training. Second, with this design we could show that the effect of training is not merely context (topic) dependent by including a variety of topics for each group of students. In addition, it allowed us to estimate the pure training effect over both intervention groups with a statistical random effects model.

The results of this study indicate (1) that we can induce a positive effect on the quality of ontology with a guideline-based training, and (2) that it is possible to quantify this effect with a CQ-based measurement instrument. We consider these results as a step into the right direction of training and evaluating ontology development processes. Nonetheless, we should ask why the effect size is so moderate although we invested a lot of effort in the training. We suggest improvements regarding two aspects. First of all, students have to be given more time and exercises to consolidate new knowledge in highly complex cognitive tasks. The other way such a study can be improved would be the sharpening of the quantitative evaluation methods.

The crucial step from informal competency questions phrased in natural language to formalized competency questions was described by Gruninger [GF95, UG96]. To formulate *formal* competency questions, at least some basic classes and relations of the target ontology must be known to use them in the axioms. This is a problem for ontologies which are in the initial phase of their development. Under these circumstances, the formalization of prospective requirements and evaluation criteria are not possible both demanded to objectify the later ontology evaluation. As there is no straightforward way from informal to formal competency questions, the method of competency question based evaluation can only be regarded as semi-formal [SS09].

Both the use of top-level categories and ODPs in the development of ontologies can lead to early prospectively formalized competency questions which should be subject of further research. Top-level ontologies and ODPs both provide a core of classes, relations and relational structures with a well-defined semantics, which can be used in the axiomatization of competency questions.

It is an open question which type of competency questions and which level of difficulty will most precisely measure a difference in quality for a given setting. In our experience, only few questions (1-2 CQs) per training ontology differentiated trained from untrained students. In contrast, most questions were answered uniformly by both groups. If it were possible to determine these threshold competency questions, the overall reliability and discriminatory power of the ontology evaluation could be largely improved. At present, the only feasible way to determine such CQs is by prototyping evaluation prior to the actual measurement.

5 Conclusion

In this study, we have shown an effect of a specific guideline-based training on the performance of ontology developers compared to the performance after unspecific training by an increase of about 10 % on the rate of correct competency questions. However, the training of ontology developers and their performance evaluation is a tedious task: Their performance is more dependent on pre-existing individual competencies than on the specific acquired skills after training. In addition, this study has shown the general applicability of competency questions in a formal ontology evaluation scenario.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) within the GoodOD project (Good Ontology Design), grant JA 1904/2-1 and SCHU 2515/1-1. We would like to thank our research assistants Hannah Guth, Maren Kechel, Gelord Opitz and Henri Seel for their help with the CQ evaluation.

References

[BJG$^+$13] Martin Boeker, Ludger Jansen, Niels Grewe, Johannes Röhl, Daniel Schober, Djamila Seddig-Raufie, and Stefan Schulz. Effects of Guideline-Based Training on the Quality of Formal Ontologies: A Randomized Controlled Trial. *PLoS ONE*, 8(5):e61425, May 2013.

[BSR$^+$12] Martin Boeker, Daniel Schober, Djamila Raufie, Niels Grewe, Johannes Röhl, Stefan Schulz, and Ludger Jansen. Teaching Good Biomedical Ontology Design. In Ronald Cornet and Robert Stevens, editors, *International Conference on Biomedical Ontologies (ICBO 2012), KR-MED Series, Graz, Austria July 21-25, 2012*, volume 897. CEUR, 2012.

[BSSH08] Elena Beißwanger, Stefan Schulz, Holger Stenzhorn, and Udo Hahn. BioTop: An Upper Domain Ontology for the Life Sciences - A Description of its Current Structure, Contents, and Interfaces to OBO Ontologies. *Applied Ontology*, 3(4):205–212, 2008.

[EMS$^+$11] J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, and C. Trojahn. Ontology Alignment Evaluation Initiative: six years of experience. *Journal on Data Semantics XV*, pages 158–192, 2011.

[Euz07] J. Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 348–353, 2007.

[FCF93] M. S. Fox, J. F. Chionglo, and F. G. Fadel. A common-sense model of the enterprise. In *Proceedings of the 2nd Industrial Engineering Research Conference*, volume 1, pages 425–429, 1993.

[GCC$^+$05] Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, Jos Lehmann, Rosa Gil, Francesco Bolici, and Onofiro Strignano. Ontology Evaluation and Validation. Technical report, 2005.

[GCCL06] Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, and Jos Lehmann. Modelling Ontology Evaluation and Validation. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, pages 140–154. Springer Berlin, 2006.

[GF94] M. Gruninger and M. S. Fox. The role of competency questions in enterprise engineering. In *Proceedings of the IFIP WG5*, volume 7, pages 212–221, 1994.

[GF95] Michael Gruninger and Mark S. Fox. Methodology for the Design and Evaluation of Ontologies. 1995.

[GW02] Nicola Guarino and Christopher Welty. Evaluating ontological decisions with OntoClean. *Commun. ACM*, 45(2):61–65, February 2002.

[HDG12] Robert Hoehndorf, Michel Dumontier, and Georgios V. Gkoutos. Evaluation of research in biomedical ontologies. *Briefings in Bioinformatics*, Advance Access, September 2012.

[KFG99] H. M. Kim, M. S. Fox, and M. Grüninger. An Ontology for Quality Management — Enabling Quality Problem Identification and Tracing. *BT Technology Journal*, 17(4):131–140, October 1999.

[MBG$^+$03] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. WonderWeb Deliverable D18. Ontology Library (final), 2003.

[NVB$^+$13] Fabian Neuhaus, Amanda Vizedom, Ken Baclawski, Mike Bennett, Mike Dean, Michael Denny, Michael Gruninger, Ali Hashemi, Terry Longstreth, Leo Obrst, Steve Ray, Ram Sriram, Todd Schneider, Marcela Vegetta, Matthew West, and Peter Yim. Ontology Summit 2013 Communiqué: Towards Ontology Evaluation across the Life Cycle, 2013.

[OCM$^+$07] Leo Obrst, Werner Ceusters, Inderjeet Mani, Steve Ray, and Barry Smith. The Evaluation of Oontologies. Toward Improved Semantic Interoperability. In Christopher J. O. Baker and Kei-Hoi Cheung, editors, *Semantic Web*. Springer US, Boston, MA, 2007.

[Se12] Barry Smith and et al. Basic Formal Ontology 2.0. Draft Specification and User's Guide. Technical report, 2012.

[SS09] Katharina Siorpaes and Elena Simperl. Human Intelligence in the Process of Semantic Content Creation. *World Wide Web*, 13(1-2):33–59, December 2009.

[SSM$^+$11] Filipe Santana, Daniel Schober, Zulma Medeiros, Fred Freitas, and Stefan Schulz. Ontology patterns for tabular representations of biomedical knowledge on neglected tropical diseases. *Bioinformatics*, 27(13):i349–i356, July 2011.

[SSSS01] S. Staab, R. Studer, H.-P. Schnurr, and Y. Sure. Knowledge processes and ontologies. *IEEE Intelligent Systems*, 16(1):26 – 34, February 2001.

[UG96] M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. *Knowledge engineering review*, 11(2):93–136, 1996.

[VG06] D. Vrandečić and A. Gangemi. Unit tests for ontologies. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, pages 1012–1020, 2006.

[VS07] Denny Vrandecic and York Sure. How to design better ontology metrics. In Enrico Franconi, Wolfgang May, and Michael Kifer, editors, *The Semantic Web: Research and Applications. Proceedings of the 4th European Semantic Web Conference (ESWC'07)*, volume 4519 of *Lecture Notes in Computer Science*, pages 311–325, Innsbruck, Austria, June 2007. Springer.