Formal Ontology in Information Systems
C. Eschenbach and M. Grüninger (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-923-3-92

Similarity as a Quality Indicator in Ontology Engineering

Krzysztof JANOWICZ [1], Patrick MAUÉ, Marc WILKES, Sven SCHADE, Franca SCHERER, Matthias BRAUN, Sören DUPKE, and Werner KUHN

Institute for Geoinformatics, University of Muenster, Germany
janowicz|patrick.maue|marc.wilkes|sven.schade|franca.scherer|m.braun|soeren.dupke|kuhn@uni-muenster.de

Abstract. In the last years, several methodologies for ontology engineering have been proposed. Most of these methodologies guide the engineer from a first paper draft to an implemented – mostly description logics-based – ontology. A quality assessment of how accurately the resulting ontology fits the initial conceptualization and intended application has not been proposed so far. In this paper, we investigate the role of semantic similarity as a quality indicator. Based on similarity rankings, our approach allows for a qualitative estimation whether the domain experts' initial conceptualization is reflected by the developed ontology and whether it fits the users' application area. Our approach does not propose yet another ontology engineering methodology but can be integrated into existing ones. A plug-in to the Protégé ontology editor implementing our approach is introduced and applied to a scenario from hydrology. The benefits and restrictions of similarity as a quality indicator are pointed out.

Keywords. ontology engineering, semantic similarity, quality assurance, requirements engineering, knowledge management

1. Introduction

Knowledge engineering deals with the acquisition, representation, and maintenance of knowledge-based systems. These systems offer retrieval and reasoning capabilities to support users in finding, interpreting, and reusing knowledge. The engineering of ontologies is a characteristic application of knowledge engineering, with ontologies as tools to represent the acquired knowledge. Various formal languages can be used to implement ontologies, i.e., to develop a computational representation for knowledge acquired from domain experts. Description Logics (DL), mostly used to implement ontologies on the Web, are a family of such languages with a special focus on reasoning services.

Answering the question how adequate the developed ontology captures the experts' initial conceptualizations (i.e., the intended meaning at a specific point in time) as well as the users' intended application area is a major issue in ontology engineering. Several methodologies offer support for knowledge acquisition and implementation, while tools for quality assessment suitable for both the domain experts and ontology users without

[1] Corresponding Author

a strong background in information science are missing. This paper proposes semantic similarity measurement as a potential quality indicator. Similarity measurement – originated in psychology – gained attention as a cognitive approach to information retrieval [1]. Inter-concept similarity rankings obtained using the SIM-DL similarity server [2] have been compared with human similarity rankings. Both correlate positively and significantly, if the natural language descriptions underlying the DL concepts were shown to the participants beforehand [3]. We therefore claim that a correlation between similarity rankings obtained from experts and the computed ontology ranking indicates whether the ontology captures the experts' initial conceptualization (given that the developed ontology was implements using the experts' input).

The paper is structured as follows. It starts with an introduction into relevant aspects of knowledge engineering (section 2) and semantic similarity measurement (section 3). Next, section 4 discusses the role of similarity as a quality indicator within the ontology engineering process. The proposed approach is applied to a hydrology use case involving existing ontologies (section 5). The benefits and restrictions of our methodology are elucidated. Finally, in section 6, conclusions and directions of further work are given.

2. Quality Assurance in Ontology Engineering

Ontologies are typically used for data annotation and integration, or to ensure interoperability between software components. In Ontology Driven Architectures [4], ontologies are included at different stages of the software engineering process. A systematic approach for the development of such ontologies is required to ensure quality. Various methodologies have been developed to accomplish a controlled and traceable engineering process. Overviews of these methodologies are given in [5]. One of the most frequently applied methodologies is *Methontology* [5].

According to *Methontology*, the ontology development process can be divided into five phases: *Specification* includes the identification of intended use, scope, and the required expressiveness of the underlying representation language. In the next phase (*conceptualization*), the knowledge of the domain of interest is structured. During *formalization*, the conceptual model, i.e., the result from the conceptualization phase, is transformed into a formal model. The ontology is implemented in the next phase (*implementation*). Finally, *maintenance* involves regular updates to correct or enhance the ontologies. This paper focuses on two activities involved in this process: *knowledge acquisition* and *evaluation*. Both are discussed in detail in the following subsections.

As illustrated in figure 1, three types of actors are involved in the development of ontologies. The steps 1-4 and the involved actions are described in section 4.

1. *Ontology users* define the application-specific needs for the ontology and evaluate whether the engineers' implementation matches their requirements.
2. *Domain experts* contribute to and agree on the knowledge which should be implemented in the ontology.
3. *Ontology engineers* analyze whether existing ontologies satisfy the experts' needs or implement the experts' conceptualization as a new ontology.

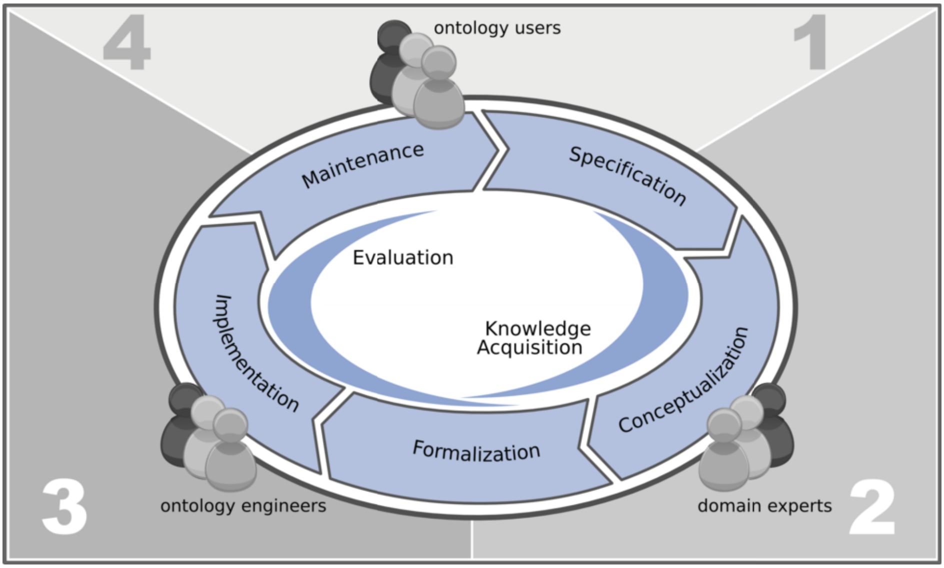

Figure 1. Phases and activities of Methontology and their relation to the actors (modified from [5]).

2.1. Knowledge Acquisition

Knowledge acquisition already starts in the specification phase, and is essential during the conceptualization phase [6]. Similar to software engineering, ontology engineering involves the identification of requirements, e.g., by specifying usage scenarios with the client. The ontology engineer is then responsible for the subsequent implementation.

Two methods can be joined to initiate knowledge acquisition. The 20-question technique [6] is a game-like approach where two persons perform a semi-structured interview. An ontology engineer (as interviewer) has a particular concept in mind, and a domain expert has to guess the concept by asking up to 20 questions. These questions have to be answered with *Yes* or *No*. All questions and answers are written down in protocols. This approach has proven to reveal concepts and relations that are central to the experts' domain [6]. Several groups need to perform these interviews to ensure suitable results. Applying the 20-question technique multiple times per group results in a rich set of protocols used as input for subsequent steps. All concepts which can directly be extracted from these protocols are used as starting point for a card sorting technique [7] . The domain experts structure a set of cards, where each card represents a concept. Without any further input they are free to order these cards. In addition, they are allowed to remove and to add cards. Building and naming clusters among the cards sketch the domain view. Using the 20-question technique in conjunction with card sorting results in a set of concepts, which can be used to generate small- to medium-sized ontologies.

Additionally, the repertory grid technique [8] can be embedded into the knowledge acquisition. In this interview technique a person compares concepts and reasons, based on their properties, why some concepts are similar while others are different. This reasoning gives information about the way a person constructs concepts. Therefore, it offers an individual domain view of the conceptualization and answers the question why the concepts are constructed in a certain way.

2.2. Evaluation

Before ontologies can be released and deployed in applications, the ontology engineers have to ensure that they meet the pre-defined quality standards. An evaluation is performed in order to validate a certain ontology according to the application-specific criteria [9], which can be further divided into technological, structural, conceptual, functional, and user-oriented aspects [10].

Functional parameters, which are related to the intended use of an ontology, are addressed by the proposed similarity-based approach. They indicate if the formalized knowledge suits the intended purpose, and if the used formalization matches the desired application. Accordingly, this facet of an ontology's quality is called *fitness for purpose* within this paper. Other parameters to assess fitness for purpose also include consistency, spelling of terms, and meeting of competency questions based on usage scenarios [11, 12].

3. Semantic Similarity Measurement

Similarity originated in psychology to investigate how entities are grouped into categories, and why some categories (and their members) are comparable while others are not [13, 14]. Similarity gained attention within the last years in computer science and especially in research on artificial intelligence [1]. In contrast to a purely structural comparison, *semantic* similarity measures the proximity of meanings. While semantic similarity can be measured on the level of individuals, concepts, or ontologies, we focus on inter-concept similarity within this paper. In dependence of the (computational) characteristics of the representation language, concepts are specified as unstructured bags of features [15], dimensions in a multi-dimensional space [16], or set-restrictions specified using various kinds of description logics [2, 17, 18, 19]. Besides applications in information retrieval, similarity measures have also been used for ontology mapping and alignment [20, 21]. As the computational concepts are models of concepts in human minds, similarity depends on what is said (in terms of representation) about these concepts.

While the proposed ontology evaluation approach is independent from a particular similarity theory, we focus on the SIM-DL [2] theory here. It has been implemented as a description logics interface (DIG) compliant semantic similarity server. In addition, a plug-in to the Protégé ontology editor has been developed to support engineers during similarity reasoning. The current release[2] supports subsumption and similarity reasoning up to the description logic $\mathcal{ALCHQ}$, as well as the computation of the *most specific concept* and *least common subsumer* up to $\mathcal{ALE}$. A human participants test (carried out using SIM-DL and the *FTO* hydrology test ontology also used within this paper) has proven that the SIM-DL similarity rankings are positively and significantly correlated with human similarity judgments [3].

SIM-DL, which can be seen as an extension of the measure proposed by Borgida et al. [19], is a non-symmetric and context-aware similarity measure for information retrieval. It compares a *search concept* C_s with a set of *target concepts* $\{C_{t_1}, ..., C_{t_m}\}$ from an ontology (or several ontologies using a shared top-level ontology). The concepts can be specified using various kinds of expressive DL. The target concepts can either

[2]The release can be downloaded at http://sim-dl.sourceforge.net/. SIM-DL is free and open source software.

be selected by hand, or derived from the *context of discourse* $\mathcal{C}_d$ [22] which is defined as the set of concepts which are subsumed by the context concept C_c ($\mathcal{C}_d = \{C_t | C_t \sqsubseteq C_c\}$). Hence, each (named) concept $C_t \in \mathcal{C}_d$ is a target concept for which the similarity $sim(C_s, C_t)$ is computed. Besides cutting out the set of compared concepts, $\mathcal{C}_d$ also influences the resulting similarities (see [2, 22] for details).

SIM-DL compares two DL concepts in canonical form by measuring the degree of overlap between their definitions. A high level of overlap indicates a high similarity and vice versa. Hence, also disjoint concepts can be similar. DL concepts are specified by applying language constructors, such as intersection or existential quantification, to primitive concepts and roles. Consequently, similarity is defined as a polymorphic, binary, and real-valued function $C_s \times C_t \rightarrow$ R[0,1] providing implementations for all language constructs offered by the used logic. The overall similarity between concepts is the normalized sum of the similarities calculated for all parts (i.e., subconcepts and superconcepts, respectively) of the concept definitions. A similarity value of 1 indicates that the compared concepts cannot be differentiated, whereas 0 shows that they are not similar at all. As SIM-DL is a non-symmetric measure, the similarity $sim(C_s, C_t)$ is not necessarily equal to $sim(C_t, C_s)$. Therefore, the comparison of two concepts does not only depend on their descriptors, but also on the direction in which both are compared.

A single similarity value (e.g., 0.67) for $sim(C_s, C_t)$ does not answer the question whether there are more or less similar target concepts in the examined ontology. It is not sufficient to know that possible similarity values range from 0 to 1 as long as their distribution is unclear. Consequently, SIM-DL delivers similarity rankings SR. The result of a similarity query is an ordered list with descending similarity values $sim(C_s, C_{t_i})$. The SIM-DL similarity server and plug-in also offer additional result representations which are more accessible for domain experts and users. These include font-size scaling (as known from tag-clouds) or the categorization of target concepts with respect to their similarity to C_s [22].

4. Similarity as Quality Measure in Ontology Engineering

This section introduces semantic similarity as a potential quality indicator. Similarity measurement does not cover all aspects of quality assurance, but rather suggests whether an ontology reflects the domain experts' initial conceptualization and the users' intended application. Consequently, semantic similarity is a candidate for assessing fitness for purpose in ontology engineering. This section describes how the ontology engineering process benefits from the proposed similarity-based approach, and how and where the three types of actors are involved. Figure 2 shows the role of similarity at certain steps of this process.

4.1. Ontology Users: Request

The ontology users request ontologies for a particular domain or application. The ontology engineering life cycle starts (step 1) and ends (step 4) with the user. In both cases the users' task is to evaluate if the available ontology fits the specific purpose, e.g., if it can be deployed in the users' application. In step 1, the ontology users have identified the need for an ontology, and therefore initiate the ontology engineering process by forwarding the request to domain experts.

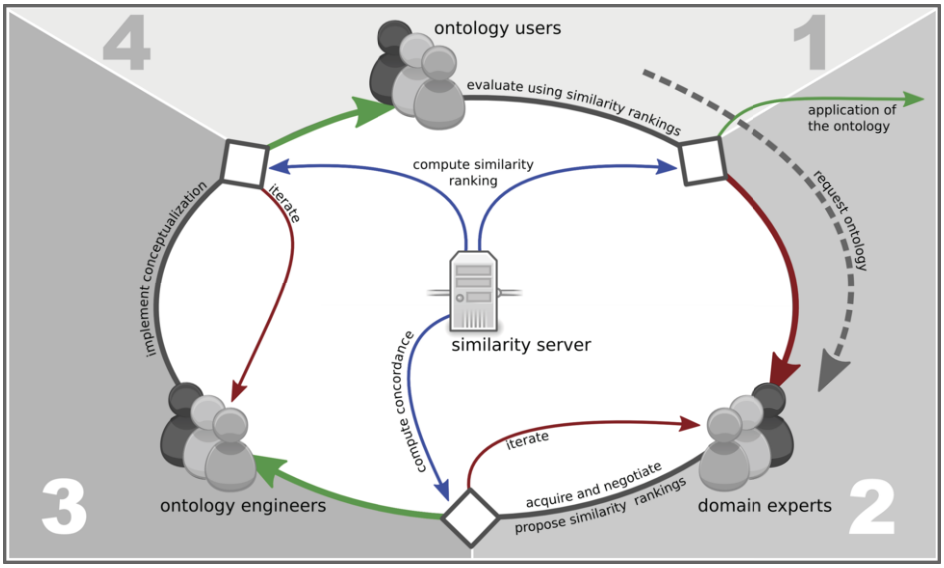

Figure 2. The role of similarity within the ontology engineering process.

4.2. Domain Experts: Knowledge Acquisition and Negotiation

Knowledge acquisition usually depends on domain experts as knowledge sources. Once they receive the users' request (step 1), the domain experts' task is to identify the requirements together with the users. The identification of scope and core concepts [5] is part of the requirements engineering process. We suggest to extend this task with the identification of the search concept C_s and a set of target concepts as well as the creation of similarity rankings SR_{de} between those concepts. These rankings will then be used as an indicator for the quality of an ontology in terms of fitness for purpose.

Current results from the SWING project [23] show that the combination of the 20-question technique and the card sorting method (see section 2.1) provide a way to identify search and target concepts. Five experts from the geology and quarrying domain participated in a knowledge acquisition process. One goal of the meeting was drafting an ontology about the transportation of aggregates. Each of the domain experts is interviewed by an ontology engineer. All concepts appearing in the protocols are used for the card sorting technique. By structuring the cards the domain experts jointly built clusters. One cluster, named "vehicle", contained the concepts Car, $Truck$, $Train$, $Bicycle$, $Pipeline$, $Boat$, and $Plane$. Another cluster ("transportation network") was built by the concepts $Motorway$, $Railroad$, $WaterCourse$, $River$, $Canal$, $Highway$, $Road$, and $Street$. The similarity-based approach to ontology evaluation can now be applied per cluster, i.e., all concepts within a cluster are potential search and target concepts. The concepts appearing most frequently in the 20-question protocols are likely to be most central for the domain. Those are chosen as search concepts, all remaining concepts of a cluster become target concepts. For the "vehicle" cluster this means $Truck$ is selected as C_s and all other concepts of the cluster make up the set of target concepts. $Road$ is selected as the search concept in the "transportation network" cluster.

The approach to identify the core concepts for the similarity rankings, and in particular C_s, is not necessarily crucial for the similarity computation. But we assume that the search concept as well as the target concepts are carefully selected and match the scope of the required ontology. As similarity rankings can be calculated on concepts from several clusters, matching the scope of even large ontologies can be fulfilled. All domain experts propose their individual similarity ranking SR_{de} with regard to the ontology's application area (step 2 in figure 2) using the identified search concept and target concepts. Next, the concordance as measure of the level of agreement between the domain experts' similarity rankings is calculated. A high (and significant) value indicates a common understanding of the core concepts by the domain experts. If the concordance is statistically insignificant (with respect to a pre-defined significance level) for the application, the domain experts' understanding of the core concepts needs to be revised (iteration at step 2). The discussion needs to clarify the definitions of each concept regarding its important characteristics. Afterwards, each domain expert performs a new similarity ranking and the concordance of these new rankings is calculated. Step 2 is repeated until a significant concordance between the similarity rankings is reached.

4.3. Ontology Engineers: Implementing the Experts' Conceptualization

Once there is a significant concordance between the similarity rankings of the domain experts, the information necessary to implement the experts' conceptualization is passed to the ontology engineers (this includes the protocols from the techniques introduced in section 2.1). In addition, an averaged similarity ranking SR_{de} is computed out of the experts' individual similarity rankings. This ranking becomes part of the requirements for the ontology. After the ontology has been developed, the ranking acts as a reference to determine whether the new ontology reflects the domain experts' initial conceptualization. Thus, the averaged ranking is used to evaluate fitness for purpose. The engineers compute a similarity ranking SR_{oe} using the SIM-DL similarity server and Protégé plug-in (see section 3 and figure 4) for the same search and target concepts as used by the domain experts. A significant and positive correlation between the domain experts' and SIM-DL's rankings indicates that the developed ontology reflects the experts' initial conceptualization. In this case, the ontology can be passed to the ontology users for further evaluation (again, using the proposed similarity ranking approach as depicted in step 4 of figure 2). If the similarity rankings do not correlate (or the correlation does not reach the pre-defined strength and significance level), an iteration in the ontology engineering process becomes necessary, i.e., step 3 is repeated until the ontology reflects the domain experts' conceptualization. If, after several iterations, no significant correlation is achieved, it might be necessary to return to the specification phase (step 2) to ensure that all relevant information from this phase is available to the engineers.

Instead of developing a new ontology, the engineers can also decide to investigate an existing ontology beforehand. In this case, the SIM-DL similarity ranking is computed using this ontology and compared to the averaged expert ranking. This requires that the external ontology uses the same concept names, else the engineers have to decide whether other names used in the external ontology can be treated as synonyms for the concepts selected by the experts. Finding synonyms may also benefit from similarity measurement, which is not discussed here but left for further work.

4.4. Ontology Users: Application

After passing all steps of the engineering process, the developed ontology is ready to be deployed. Following figure 1, the ontology users are also involved in the maintenance of the ontology. Up to now, the computed similarity ranking SR_{oe} and the averaged similarity ranking SR_{de} provided by the domain experts are available. But even the best correlation between these two rankings does not necessarily mean that the ontology match the users' view. With the last missing similarity ranking SR_{ou}, we compute the correlation between the rankings SR_{oe} from the engineered ontology and those from the users (step 4). SR_{ou} is also an averaged similarity ranking collected from the ontology users during the maintenance phase, e.g., using questionnaires or user feedback techniques built into the software. The knowledge and therefore also the conceptualization of a particular domain can evolve over time, which means this step has to be performed regularly.

If a significant correlation between SR_{ou} and SR_{oe} exists and does not change over time, it can be assumed that the ontology represents the users' view with respect to the application. A low correlation between SR_{ou} and SR_{oe} might imply that the ontology does, in its current state, not satisfy the users' needs. Re-initiating the ontology engineering life cycle, including the users' similarity rankings, is advisable.

5. Application

This section applies the steps described in section 4 to a set of concepts from four different ontologies to demonstrate our approach. The similarity between these concepts is measured and the resulting ranking is compared to a similarity ranking defined by the authors of this paper acting as domain experts and users, respectively. The concepts and ontologies were chosen to elucidate selected aspects of similarity as a quality indicator. An evaluation involving external domain experts and ontology engineers is left for further work. The used ontologies are excerpts of the hydrology ontology from Ordnance Survey *OS Hydrology*[3], a (OWL-Lite) version of the Alexandria Digital Gazetteer Feature Type Thesaurus *ADL FTT*[4], the *AKTiveSA* ontology[5], and the Feature Type Ontology *FTO Hydrology*[6] developed by the authors for the human participants test described by Janowicz et al. [3]. Figure 3 gives a brief overview over the hydrological concepts within these ontologies; interested readers are referred to the online OWL versions.

In the following we assume that users of a specific hydrology application such as a decision support system for an agency request an ontology. Domain experts analyze the users' requirements and identify core concepts for the new hydrology ontology using the 20-question and card sorting technique. The resulting core concepts are *Canal*, as search concept, and *River*, *Lake*, *IrrigationCanal*, *Ocean*, *Reservoir*, and *OffshorePlatform* as target concepts.

After deciding on the core concepts, and negotiation how these concepts should be specified, each domain expert defines a similarity ranking to express her initial conceptualization. All rankings are performed independently and afterwards compared for con-

[3]http://www.ordnancesurvey.co.uk/oswebsite/ontology/
[4]http://ifgi.uni-muenster.de/ janowicz/downloads/FTT-v01.owl
[5]http://www.edefence.org/ ps/aktivesa/OntoWeb/index.htm
[6]http://sim-dl.sourceforge.net/downloads/

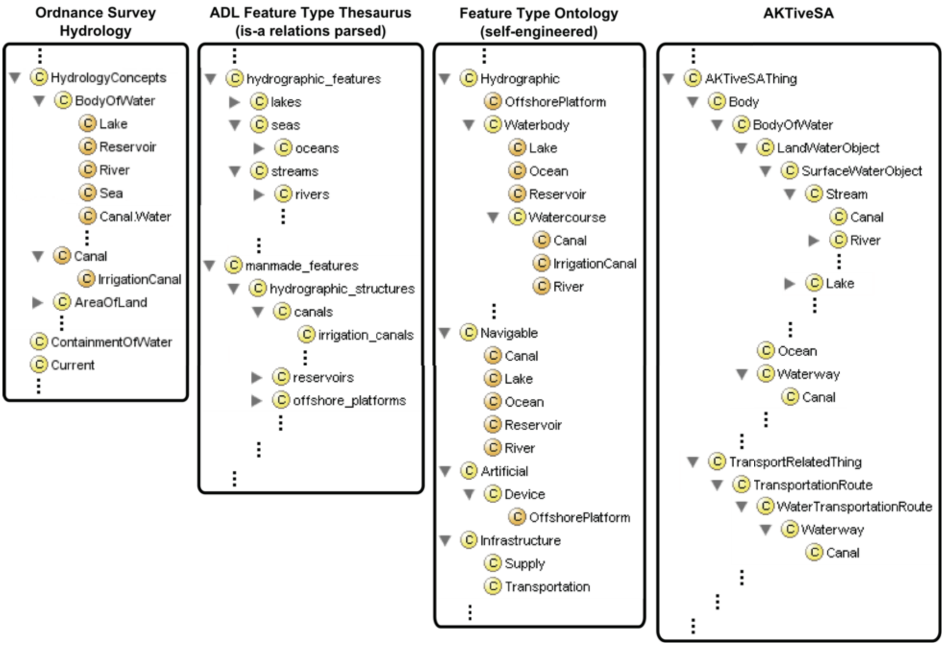

Figure 3. Overview of the four ontologies used for similarity measurement.

cordance using Kendall's coefficient of concordance W as measure of the level of agreement between the domain experts. In case of the authors' rankings this yields $W = 0.77$, which is a statistically significant result (using a significance level of 0.05).

The averaged similarity ranking by the domain experts is passed on to both the users and the ontology engineers. The users might refine the requirements if the domain experts' rankings do not match the users' expectations. The ontology engineers use these rankings for later verification of the implemented ontologies. The computed similarity rankings are then compared with those produced by the domain experts.

To measure similarity and compare the resulting rankings for correlation, the SIM-DL similarity server is used in conjunction with an extended version of the Protégé similarity plug-in. As depicted in figure 4 the extension offers a tab for estimating the similarity between the search and the target concepts using sliders. The resulting ranking and the similarity values are compared to the results obtained from the SIM-DL server.

The Protégé extension shown in figure 4 not only allows for specifying a ranking of concepts, but also for expressing a quantitative distance between these concepts. However, different people (i.e., domain experts) use different (cognitive) similarity scales and distributions [3]. Hence, the interpretation of the absolute similarity values and distances between them is difficult. Consequently, this paper focuses on similarity rankings.

The *FTO Hydrology* ontology is supposed to be the ontology developed by the ontology engineers based on the experts' conceptualization. Figure 4 shows the resulting chart and correlation based on the averaged similarity ranking of the experts and the results computed by SIM-DL for the *FTO Hydrology* ontology. As shown in table 1, there is a positive ($r_s = 0.94$) and significant ($p < 0.05$) correlation between both rankings. These

Table 1. Similarity rankings for the used ontologies with respect to *Canal* as search concept.

Similarity ranking to Canal	River	Irr. Canal	Reservoir	Lake	Ocean	Off. Platform	Correlation$\diamond$
Experts' Ranking	1	2	4	3	5	6	—
ADL FTT Ranking	3	1	2	3	3	2	0.06
OS Hydrology Ranking	3	1	4	2	4$\star$	—	0.67
FTO Hydrology Ranking	1	2	3	4	5	6	0.94
AKTiveSA Ranking	1	—	2	2	3	—	0.95

$\diamond$: Spearman's rank correlation r_s measured to the experts' averaged ranking.

$\star$: The concept *Sea* is used as no concept named *Ocean* is available in the ontology.

results indicate that the *FTO Hydrology* ontology reflects the experts' conceptualization. The ontology is then passed to the users for further evaluation.

The users evaluate the received ontology using their similarity rankings in order to investigate if the ontology can be deployed into the final hydrology application. Otherwise, the users can initiate a new iteration cycle starting again with the domain experts.

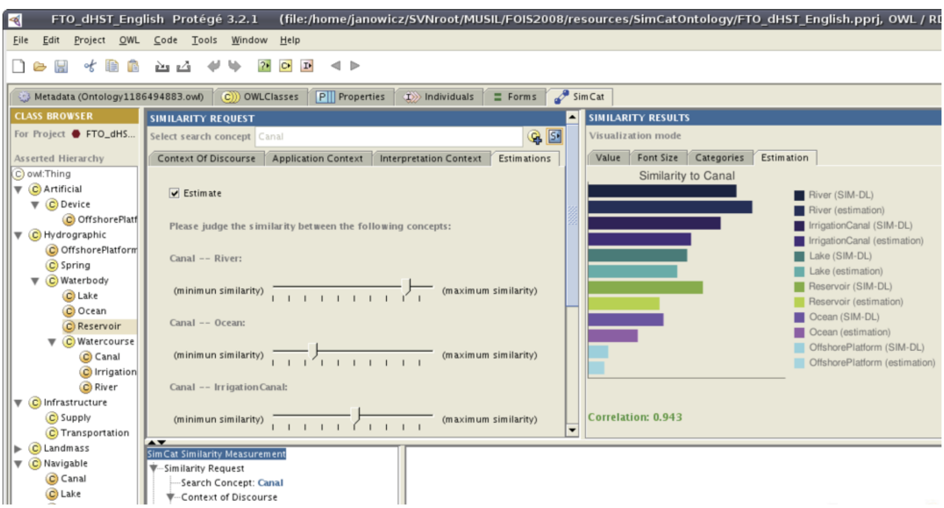

Figure 4. The extended SIM-DL Protégé plug-in with the new estimation tab (compare to [2, 22]).

It is reasonable to assume that ontology engineers first check for existing external ontologies before developing a new one. They compare the SR_{de} ranking of the experts with those from the external ontologies (in our case the *ADL FTT*, *OS Hydrology*, and *AKTiveSA* ontologies) to evaluate their fitness for purpose.

Unlike the self-engineered *FTO Hydrology* ontology, table 1 shows that the ranking obtained from the *ADL FTT* ontology does not correlate with the experts' ranking. For instance, the *ADL FTT* concept *offshore platforms* is ranked in second place, and hence above *rivers*. This can be explained with the single-inheritance structure used within this ontology, i.e., a concept cannot be a direct subconcept of two different concepts. As a consequence, the top-level distinction between *hydrographic features* and *manmade features*, and the definition of the concept *hydrographic structures* as a subclass of *manmade features*, implies that all concepts classified as *hydrographic structures* are considered manmade, but not *hydrographic features* (see figure 3). As the *ADL FTT* ontology was derived from automatically parsing the thesaurus, the subsumption relationship is the only one which could be used to measure conceptual overlap (and hence similarity). Consequently, the similarity between concepts which are not beneath a common superconcept (such as *canals* and *rivers*) is low. In contrast, sharing the same superconcept increases similarity as for *canals* and *offshore platforms*. Both are *hydrographic structures*[7] and *manmade features*. Such view does not reflect the experts' initial conceptualization, and therefore the ontology cannot be used for the hydrology application (or requires substantial modification).

A test run for the second external ontology, an excerpt from *OS Hydrology*, shows a positive ($r_s = 0.67$) but insignificant correlation to the experts' ranking. This is due to several reasons: first, the concepts *OffshorePlatform* and *Ocean* are not part of this ontology which decreases the number of ranked concepts decisively. Second, the implemented concepts do not meet the experts' conceptualization. As described in section 4.3 the *OS Hydrology* concept *Sea* is chosen as potential alternative for *Ocean* within this example. The surprising result that *Lake* is more similar to *Canal* than *River* can be explained as follows. First, while *River*, *Lake*, *Sea*, and *Reservoir* are subconcepts of *BodyOfWater*, *Canal* and *IrrigationCanal* are not (see figure 3). However, there is a subconcept of *BodyOfWater* called *Canal.Water* that comprises some of the intended characteristics missing in *Canal* (e.g., being navigable). Second, in contrast to *Canal* and *Lake*, the definition of *River* does not contain a value restriction for being connected to other bodies of water.

The *AKTiveSA* ontology represents the case where a high correlation ($r_s = 0.95$) indicates that the concepts reflect the experts' conceptualization. However, not all concepts are defined in the ontology, and hence the correlation is statistically insignificant. No candidate concepts for *OffshorePlatform* and *IrrigationCanal* were found. In this case, the engineers can decide to extend the ontology with the missing concepts and recalculate the correlation.

Summing up, the application of similarity as quality indicator points to the following benefits and shortcomings. Similarity helps to assess if developed ontologies reflect the intended conceptualizations of experts and users. Simplicity is a desired prerequisite for an evaluation method in order to be adopted by non-technical experts. As similarity is grounded in cognition, the cognitive effort imposed on actors to produce similarity rankings is low. This is especially important for non-technical domain experts and end-users lacking formal background on description logics. Therefore, similarity rankings provide the engineer with the possibility to integrate the users and experts during the implementation phase. SIM-DL compares concepts for overlapping definitions. This does not guarantee that these definitions are relevant for the particular application. For an external ontology this may cause a correlating similarity ranking, although the definitions focus on other applications (such as recreation instead of navigation). Therefore, SIM-DL allows to set the context of discourse (see section 4.3) to enforce particular concept definitions. Finally, similarity does not answer the question how concepts differ. To improve the expressivity of similarity as quality indicator, it should therefore be combined with *difference* operations as proposed by Teege [24].

[7] Which is surprising as the thesaurus defines hydrographic structure as "constructed bodies of water".

6. Conclusions and Further Work

Ontology engineering and similarity reasoning have only been remote cousins so far. We have shown in this paper that semantic similarity rankings founded in formal ontology can support the ontology engineering process. In particular, they serve as measures for how accurately an ontology matches the conceptualizations held by ontology engineers and users. Our approach is orthogonal to ontology engineering methods and can be incorporated into any of them. The contributed plug-in to the Protégé ontology editor serves this purpose and has been successfully tested in a scenario with hydrological information. While we focused on the simplified hydrology example here, a more sophisticated scenario from quarry mining involving external domain experts and users is under development in the SWING project (see section 4.2).

Our main contribution is toward the problem of quality assurance for information system ontologies. The simple idea to compare similarity rankings of concept specifications in natural language (produced by domain experts or users) with those of concept specifications in DL (produced by ontology engineers) represents an effective way of assessing how closely the stated constraints on meaning match the intended meaning.

Our method is rooted in formal ontology, as the semantic similarity rankings are based on a similarity theory that accounts for concept specifications instead of a purely syntactical measure. The similarity theory and its application have been developed with theoretical foundations in psychological literature on similarity and the logics to express them. All similarity measures crucially depend on the representation chosen for the compared concepts. A solid grounding in formal ontology can therefore be expected to improve the match between human and computational similarity rankings. This has been shown to be the case by Janowicz [3]. In this paper, we have used a non-symmetric similarity measure. SIM-DL also supports symmetric similarity; further work should investigate which approach fits better for quality assessment.

Beyond the formal foundations, the iterative engineering model involving three actors (domain expert, knowledge engineer, user) represents a way toward more realistic knowledge acquisition and management scenarios. The social nature of these processes, particularly the fact that specifications of conceptualizations are negotiated among the participants, is ideally supported by a concise and transparent quality measure (which is also easy to use) such as the match between similarity rankings.

From a formal ontology point of view, a benefit of our approach is that it can reveal incomplete concept definitions. For instance, in *AKTiveSA*, canals differ from other bodies of water by also being transportation routes. Length is a characteristic of transportation routes, but not automatically of rivers, since it does not apply to all bodies of water. The lacking length of rivers has a negative impact on the similarity value of *Canal* to *River*. It indicates to the ontology engineer that, from a certain perspective, the ontology is incomplete or inhomogeneous.

7. Acknowledgements

The comments from three anonymous reviewers provided useful suggestions to improve the content and clarity of the paper. This work is funded by the SimCat project granted by the German Research Foundation (DFG Ra1062/2-1) and the SWING project which is funded by the European Commission (FP6-26514). We also acknowledge the contribution of five domain experts from the French geological survey, BRGM (Bureau de Recherches Géologiques et Minières).

References

[1] E. L. Rissland. AI and similarity. *IEEE Intelligent Systems*, 21(3):39–49, 2006.

[2] K. Janowicz, C. Keßler, M. Schwarz, M. Wilkes, I. Panov, M. Espeter, and B. Baeumer. Algorithm, Implementation and Application of the SIM-DL Similarity Server. In *Second International Conference on GeoSpatial Semantics (GeoS 2007)*, number 4853 in Lecture Notes in Computer Science, pages 128–145. Springer, 2007.

[3] K. Janowicz, C. Keßler, I. Panov, M. Wilkes, M. Espeter, and M. Schwarz. A Study on the Cognitive Plausibility of SIM-DL Similarity Rankings for Geographic Feature Types. In *11th AGILE International Conference on Geographic Information Science (AGILE 2008)*, Lecture Notes in Computer Science, Girona, Spain, 5-8. May 2008; forthcoming. Springer.

[4] H. Happel and S. Seedorf. Applications of ontologies in software engineering. In *International Workshop on Semantic Web Enabled Software Engineering (SWESE'06)*, Athens, USA, November 2006.

[5] A. Gómez-Pérez, O. Corcho, and M. Fernández-López. *Ontological Engineering: with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web.* Springer, July 2004.

[6] S. Schade, P. Maué, J. Langlois, and E. Klien. Ontology engineering with domain experts - a field report. In *European Geosciences Union - General Assembly*, 2008.

[7] N. J. Cooke. Varieties of knowledge elicitation techniques. *Int. J. Hum.-Comput. Stud.*, 41(6):801–849, 1994.

[8] D. Jankowicz. *Easy Guide to Repertory Grids*. Wiley, London., 2004.

[9] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann. A theoretical framework for ontology evaluation and validation. In *Semantic Web Applications and Perspectives (SWAP) – 2nd Italian Semantic Web Workshop*, 2005.

[10] S. Schade, E. Klien, P. Maué, D. Fitzner, and W. Kuhn. Ontologies in the SWING-Project: Report on Modelling Approach and Guideline - Deliverable 3.2, 2008.

[11] M. Grüninger and M. S. Fox. Methodology for the design and evaluation of ontologies. In *Proceedings of IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.

[12] M. Uschold and M. Grüninger. Ontologies: Principles, methods and applications. *Knowledge Engineering Review*, 10(2):92–155, 1996.

[13] R. L. Goldstone and J. Son. Similarity. In K. Holyoak and R. Morrison, editors, *Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press, 2005.

[14] D. Medin, R. Goldstone, and D. Gentner. Respects for similarity. *Psychological Review*, 100(2):254–278, 1993.

[15] A. Rodríguez and M. Egenhofer. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, 18(3):229–256, 2004.

[16] M. Raubal. Formalizing conceptual spaces. In A. Varzi and L. Vieu, editors, *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, volume 114 of *Frontiers in Artificial Intelligence and Applications*, pages 153–164. IOS Press, Amsterdam, NL, 2004.

[17] C. d'Amato, N. Fanizzi, and F. Esposito. A semantic similarity measure for expressive description logics. In *CILC 2005, Convegno Italiano di Logica Computazionale*, Rome, Italy, 2005.

[18] R. Araújo and H. S. Pinto. Semilarity: Towards a model-driven approach to similarity. In *International Workshop on Description Logics*, volume 20, pages 155–162. Bolzano University Press, June 2007.

[19] A. Borgida, T. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In *Proceedings of the 2005 International Workshop on Description Logics (DL2005)*, volume 147 of *CEUR Workshop Proceedings*. CEUR, Edinburgh, Scotland, UK, 2005.

[20] A. Maedche and S. Staab. Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, number 2473 in Lecture Notes in Computer Science, pages 251 – 263. Springer, 2002.

[21] W. Sunna and I. F. Cruz. Structure-based methods to enhance geospatial ontology alignment. In *GeoSpatial Semantics, Second International Conference, GeoS 2007*, number 4853 in Lecture Notes in Computer Science, pages 82–97. Springer, 2007.

[22] K. Janowicz. Kinds of contexts and their impact on semantic similarity measurement. In *5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea) at the 6th IEEE International Conference on Pervasive Computing and Communication (PerCom'08)*, Hong Kong, March 2008. IEEE Computer Society.

[23] D. Roman and E. Klien. *The Geospatial Web. In the Advanced Information and Knowledge Processing Series*, chapter SWING - A Semantic Framework for Geospatial Services, pages 229–234. Springer, 2007.

[24] G. Teege. Making the difference: A subtraction operation for description logics. In J. Doyle, E. Sandewall, and P. Torasso, editors, *4th International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 540–550, Bonn, Germany, 1994. Morgan Kaufmann.