KRZYSZTOF JANOWICZ

COMPUTING SEMANTIC SIMILARITY AMONG GEOGRAPHIC FEATURE TYPES REPRESENTED IN EXPRESSIVE DESCRIPTION LOGICS

Juni 2008

Geoinformatik

COMPUTING SEMANTIC SIMILARITY AMONG GEOGRAPHIC FEATURE TYPES REPRESENTED IN EXPRESSIVE DESCRIPTION LOGICS

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Naturwissenschaften
im Fachbereich Geowissenschaften
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von
Krzysztof Janowicz
aus Poznań, Polen
im Juni 2008

Dekan:	Prof. Dr. Hans Kerp
Erstgutachter:	Prof. Dr. Werner Kuhn
Zweitgutachter:	Prof. Dr. Martin Raubal
Tag der mündlichen Prüfung:	....
Tag der Promotion:	....

ABSTRACT

Similarity measurement gained attention within semantics-based information retrieval over the last years. Based on research from psychology and information science, several theories investigate how to compute the similarity between entities, concepts, or spatial scenes. However, real implementations and applications are still rare. Most existing similarity theories use their own representation language while the majority of (geo-)ontologies is annotated using various kinds of description logics. One reason for this gap is that most theories cannot handle the expressivity of description logics. In addition, the interpretation of similarity values is not trivial, e.g., a high similarity between *River* and *Canal* does not necessarily mean that canals could be used as replacements for rivers within a particular scenario (e.g., recreation).

This thesis introduces a generic framework to explain how similarity theories are used for information retrieval and what they measure. Based on this framework, the context-aware SIM-DL theory is introduced to close the gap between description logics-based ontologies and similarity theories. A DIG (DL Implementation Group) interface compliant semantic similarity server and an extension to the Protégé ontology editor are introduced. The impact of similarity for semantics-based information retrieval is emphasized by use cases from gazetteer research, including a vision of a distributed gazetteer infrastructure and a similarity-enabled gazetteer Web interface. Two human participants tests are carried out to verify whether similarity rankings computed using SIM-DL correlate with human similarity estimations. An extended context model for similarity is presented and the interpretation of similarity values is discussed. Further directions of research are pointed out. This includes first ideas on how to improve the task-dependent selection of relevant features, as well as various optimization techniques, such as approximation, to reduce the computational cost of similarity measurement.

PUBLICATIONS

Several ideas, fragments, and figures have appeared previously in the following publications:

K. Janowicz. Extending semantic similarity measurement with thematic roles. In A. M. Rodríguez, I. F. Cruz, M. Egenhofer, and S. Levashkin, editors, First International Conference on GeoSpatial Semantics (GeoS '05), volume 3799 of Lecture Notes in Computer Science, pages 137–152. Springer, 2005.

K. Janowicz. Towards a similarity-based identity assumption service for historical places. In M. Raubal, H. J. Miller, A. U. Frank, and M. F. Goodchild, editors, Geographic Information Science, 4th International Conference (GIScience 2006), volume 4197 of Lecture Notes in Computer Science, pages 199–216, September 2006.

K. Janowicz. SIM-DL: Towards a semantic similarity measurement theory for the description logic ALCNR in geographic information retrieval. In R. Meersan, Z. Tari, and P. Herrero, editors, On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, OTM Confederated International Workshops and Posters, AWeSOMe, CAMS, COMINF, IS, KSinBIT, MIOS-CIAO, MONET, OnToContent, ORM, PerSys, OTM Academy Doctoral Consortium, RDDS, SWWS, and SeBGIS. Proceedings, Part II, volume 4278 of Lecture Notes in Computer Science, pages 1681 – 1692. Springer, Montpellier, France, October 2006.

K. Janowicz. Similarity-based retrieval for geospatial semantic web services specified using the web service modeling language (wsml-core). In Terra Cognita – Directions to the Geospatial Semantic Web Workshop, Workshop at the International Semantic Web Conference 2006, Athens, USA, November 2006.

K. Janowicz. Similarity-based retrieval for geospatial semantic web services specified using the web service modeling language (wsml-core). In A. Scharl and K. Tochtermann, editors, The Geospatial Web – How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society, Lecture Notes in Computer Science, pages 235–245. Springer, Berlin, 2007.

K. Janowicz and M. Raubal. Affordance-based similarity measurement for entity types. In S. Winter, M. Duckham, L. Kulik, and B. Kuipers, editors, Spatial Information Theory, 8th International Conference, COSIT 2007, volume 4736 of Lecture Notes in Computer Science, pages 133–151. Springer, Melbourne, Australia, September 2007.

K. Janowicz, C. Keßler, M. Schwarz, M. Wilkes, I. Panov, M. Espeter, and B. Bäumer. 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.

C. Keßler, M. Raubal, and K. Janowicz. The effect of context on semantic similarity measurement. In R. Meersman, Z. Tari, and P. Herrero, editors, On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops, OTM Confederated International Workshops and Posters, AWeSOMe, CAMS, OTM Academy Doctoral Consortium, MONET, OnToContent, ORM, PerSys, PPN, RDDS, SSWS, and SWWS 2007, Proceedings, Part II, number 4806 in Lecture Notes in Computer Science, pages 1274–1284. Springer, Vilamoura, Portuga, November 2007.

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, IEEE Computer Society, March 2008.

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 L. Bernard, A. Friis-Christensen, and H. Pundt, editors, 11th AGILE International Conference on Geographic Information Science (AGILE 2008),Lecture Notes in Geoinformation and Cartography, pages 115–133. Springer, Girona, Spain, May 2008.

K. Janowicz, M. Wilkes, and M. Lutz. Similarity-based Information Retrieval and its Role within Spatial Data Infrastructures. 5th International Conference on Geographic Information Science. September 2008; forthcoming. Park City, Utah, USA.

K. Janowicz and C. Keßler. The role of ontology in improving gazetteer interaction. International Journal of Geographical Information Science, 10, 2008; forthcoming.

Each waking day is a stage dominated for good or ill,
in comedy, farce or tragedy by a dramatis persona, the 'self'.
And so it will be until the curtain drops.

Sir Charles Scott Sherrington

ACKNOWLEDGMENTS

Many friends and colleagues supported me during the last years and helped me in preparing this thesis. I would like to thank all of them for their support.

First, I want to thank my supervisors Werner Kuhn and Martin Raubal for investing so much time and effort in discussions and Ph.D. meetings, for reviewing my publications, and especially for their constructive criticism during all stages of this thesis.

I want to thank the members of the *Semantic Similarity Measurement for Role-Governed Geospatial Categories* (SimCat) project Marius Austerschulte, Boris Bäumer, Martin Espeter, Ilija Panov, and especially Mirco Schwarz, Carsten Keßler, and Marc Wilkes for co-authoring papers on semantic similarity and context, as well as for implementing various parts of the SIM-DL similarity server, Protégé plug-in, test ontology, and context model. Additionally, I would like to thank Carsten Keßler and Marc Wilkes for discussing and reading my thesis many times.

I also thank the members of the *Muenster Semantic Interoperability Lab* (MUSIL), and here especially Florian Probst. Moreover, thanks to Linda Hill, Michael Lutz, Tom Bittner, Catharina Riedemann, Antonio Krüger, Angela Schwering, Brent Hecht, and in particular Andrea Rodríguez for the fruitful discussions and comments while preparing this thesis.

Finally, I want to thank my family and friends for their encouragement and patience.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

This chapter motivates the need for a semantic similarity measure for geographic feature types represented in description logics. A research hypothesis is presented and methods to verify this hypothesis are introduced. The expected results and their impact on further work about similarity are discussed. The structure of the thesis is outlined at the end.

1.1 INTRODUCTION AND MOTIVATION

Measures of semantic similarity have a long tradition in cognitive science and psychology [56]. Similarity estimations seem to be one of the fundamental processes underlying human categorization [123]. Psychology has investigated several kinds of similarities over the last fifty years, including similarity between entities, concepts, or complex (pictorial) scenes. Different approaches to modeling similarity have been developed. While feature-based models [168] are the most prominent ones, network-based [137], geometric [165][163][151], alignment [51], and transformation-based [65] approaches have evolved over the last decades. The main motivation underlying this kind of research is to understand human cognition.

In contrast, new research from artificial intelligence (and information science in general) applies computational similarity theories as tools for information retrieval [126, 141]. Some of these approaches focus on issues of scalability to improve keyword-based search when iterating over millions of vectors [13]. Others focus on semantic similarity between concepts in ontologies [2, 3, 18, 32, 33, 79, 85, 142, 143]. The Matching Distance Similarity Measure [142, 143] was one of the first similarity measures which has been developed specifically for the geospatial domain.

In information science, similarity is especially promising with respect to user interaction. It has the potential to improve the interaction with complex search and knowledge organization systems. For instance, the ConceptVISTA[1] ontology management tool uses similarity for browsing through concepts, but also for negotiation, i.e., to establish a common agreement among domain experts. Besides subsumption reasoning, similarity is a potential candidate for semantics-based information retrieval[2] using ontologies [78, 111, 126]. Several geospatial applications are candidates for the implementation of similarity-based information retrieval techniques. Geo-portals, for example, could

1 http://www.geovista.psu.edu/ConceptVISTA

2 The term *semantics-based information retrieval* is a bit misleading, it is used here (in accordance with the literature) to distinguish various techniques from purely syntactic matchmaking (sometimes also called keyword-based search). All these techniques share the fact that they do not only rely on the user's input but try to take further information into account. A classical example is to take the subconcepts of a search concept into account, e.g., searching for bodies of water would also return rivers.

supplement search result pages with matches that do not exactly fit the user's query, but share certain characteristics with the search phrase or concept. Similarity supports novel user interfaces which allow for imprecise input and do not require the user to know about the internal structure of the queried resource. Location-based services could derive points of interest in the user's vicinity from similar, previously visited places. This also includes recommendation systems which propose similar activities to users or groups as described by Espeter [39]. In addition, similarity plays a prominent role in the semi-automatic alignment of geo-ontologies [160].

In general, the more information resources become available (e.g., via the semantic Web), the higher is the need for tools supporting the interaction with these resources and to establish bridges which connect local knowledge communities. Similarity is one of these tools, as it supports users in retrieving, and browsing through information and hence in knowledge acquisition[3].

1.2 PROBLEM STATEMENT

Several requirements have to be fulfilled to make use of the potential of similarity for semantics-based information retrieval.

- The similarity theory has to support the language used to describe the information to be retrieved. In case of ontologies, description logics are the most prominent representation language.
- The results, i.e., similarity values or rankings, produced by the similarity theory have to be explainable in terms of the used representation language (e.g., a particular description logic).
- The results produced by the similarity theory have to be cognitively plausible, i.e., they have to correlate with human similarity judgments.
- The similarity theory has to be implemented and made accessible via existing tools used for semantics-based information retrieval and ontology engineering (e.g., via the Protégé ontology editor).

If one of these requirements cannot be fulfilled, the similarity theory cannot develop its full potential. The majority of (semantic Web) ontologies are specified using various kinds of description logics. A similarity theory using its own proprietary representation language can only improve semantics-based information retrieval up to a limited degree. If the theory returns similarity estimates which conflict with the semantics of the underlying language, the results may (depending on the level of violation) become useless. For instance, if two concept descriptions have no overlap, the similarity has to be 0[4]; if the overlap is total, the similarity needs to be 1. A similarity theory which produces results that are in conflict with human similarity reasoning is

3 If we assume that information is the amount of available data and knowledge is the act of making sense of these data. A webpage in a foreign language is information, which needs to be translated and understood to derive knowledge of it.

4 Strictly speaking, there is an exception to this rule when comparing primitives; see chapter 5.

also of questionable value (at least for human-computer interaction). Finally, if the similarity theory cannot be implemented or integrated with existing tools or frameworks, its impact will be limited.

Up to now, there is no single similarity theory which fulfills these four requirements[5].

1.3 HYPOTHESIS

The goal of this thesis is to develop a semantic similarity theory (called SIM-DL) which fulfills the four requirements defined above. While all of these are taken into account, the hypothesis focuses on the third requirement, i.e., to develop a cognitively plausible similarity theory (which can cope with the expressivity of DL-based ontologies).

SIMILARITY RANKINGS OBTAINED BY COMPARING CONCEPTS USING THE SIM-DL THEORY CORRELATE STRONGLY WITH HUMAN SIMILARITY JUDGMENTS.

1.4 METHODS

To verify or falsify the hypothesis the following steps are taken in this thesis: First, a use case demonstrates the benefits of semantic similarity measurement for GIScience. Recent research on gazetteers is used to motivate the need for similarity. The potential impact of similarity for a distributed gazetteer infrastructure as well as for a new kind of Web interface is pointed out. Based on this use case, an ontology for geographic feature types is developed using the $\mathcal{ALCHQ}$ description logic. Fragments of this ontology are later on used to measure inter-concept similarity and to perform a human participants test. In the next step, a generic framework is developed based on shared characteristics of existing similarity measures. Based on this framework, the SIM-DL similarity theory is developed which can handle the expressivity of the $\mathcal{ALCHQ}$ description logic. For the gazetteer use case, the similarity theory focuses only on the ontological specifications to measure inter-concept similarity. No additional (ABox) knowledge about entities (geographic features) is taken into account. The resulting theory is implemented within a semantic similarity server which supports queries from existing ontology editors such as Protégé[6]. To achieve this, the server implements the DIG description logics interface [15]. A prototypical similarity plug-in for Protégé is presented[7]. Finally, a human participants test is carried out to examine the correlation between the SIM-DL similarity ranking and human similarity rankings. The consensus among participants is examined.

The hypothesis is verified if there is a *strong*, *positive*, and *significant*[8]

---

5 This thesis was started in April 2004; meanwhile some theories are moving towards achieving these design goals — the SIM-DL similarity server and theory presented in this thesis is one of them.

6 http://protege.stanford.edu/

7 Both the plug-in and server are developed within the *Semantic Similarity Measurement for Role-Governed Geospatial Categories* (*SimCat*) project granted by the German Research Foundation (DFG Ra1062/2-1).

8 We assume a significance level of $\alpha = 0.05$; see [19, p. 114].

correlation between human and machine rankings, and a significant consensus between the test participants. To demonstrate that the functions used to measure similarity in SIM-DL do not violate the underlying representation language, their derivation from set theory is shown.

1.5 RESULTS AND RELEVANCE FOR FURTHER WORK

The thesis presents a generic framework which describes how similarity is measured for tasks such as information retrieval. Up to now, there is no such framework which makes it difficult to compare similarity measures and to understand what exactly they measure. The proposed framework allows each theory to specify the semantics of similarity.

The Sim-DL theory and especially its implementation in the similarity server (and Protégé plug-in) helps to bridge the gap between knowledge representation on the Web, i.e., DL-based ontologies, and various existing similarity theories that were not able to cope with the expressivity of description logics so far.

The thesis introduces a classification of contexts and describes their impact on similarity measurement. Further research can use this classification to examine which context information is essential to achieve meaningful similarity estimates and which could be left aside (see Keßler [91] and Keßler et al. [92] for details).

We hope that this thesis (and the developed software) helps to establish similarity reasoning as a tool used by ontology engineers and for user interfaces in the same way as subsumption reasoning is used nowadays.

1.6 OUTLINE

The thesis is structured into 8 chapters followed by an appendix. Apart from chapter 1, 2, and 3 each chapter is an extended version of at least one workshop, conference, or journal publication.

Chapter 2 introduces semantic similarity measures with a focus on feature-based, alignment, and network-based theories. A framework defining the steps involved in measuring similarity is described.

Chapter 3 introduces the notion of ontologies and description logics as one possible representation language for ontologies.

Chapter 4 introduces the gazetteer use case. Besides an insight into research related to gazetteers, the vision of a distributed gazetteer infrastructure is elucidated and the role of similarity reasoning within such an infrastructure is pointed out. The need for a feature type ontology is discussed and first steps towards developing such an ontology are presented. A new kind of similarity-enabled Web interface for gazetteers is introduced. This chapter is based on [78, 83, 85].

In chapter 5, the SIM-DL similarity theory is introduced and discussed in detail. SIM-DL is compared to existing inter-concept measures by pointing out differences and commonalities. This chapter is an extended version of [77, 79, 80, 81, 85].

Chapter 6 discusses the implementation of the SIM-DL theory within a DIG-compliant semantic similarity server and the plug-in for the

Protégé ontology editor. Extensions to DIG, necessary for similarity reasoning, are presented. This chapter is based on [85, 86].

Based on the gazetteer use case and the SIM-DL similarity theory, chapter 7 introduces two human participants tests. The first test gives insights into the similarity between so-called role-filler pairs. It is designed to determine how SIM-DL should handle such expressions. The second test compares inter-concept similarity rankings from SIM-DL to those from human participants. The correlation between human and machine rankings as well as the consensus between the participants is computed. The evaluation chapter is an extended version of [81, 86].

Chapter 8 presents the conclusions and introduces directions for future work. Besides possible extensions to SIM-DL, several kinds of contexts and their impact on similarity are discussed. These contexts are a refinement of the so-called *context concept* originally defined for SIM-DL and are a reaction to the results from the human participants test. The kinds of contexts discussed in this chapter have been introduced in [82, 92]. The outlook on salient feature selection is based on [84].

The appendix gives some additional information about the implementation of the similarity functions in the similarity server. The server, plug-in, and ontology can be downloaded from the SimCat project Website at http://sim-dl.sf.net. Readers are pointed to the download links as well as additional resources within the respective chapters.

2 SEMANTIC SIMILARITY

This chapter gives an overview of semantic similarity measures with a focus on feature-based, alignment, and network-based approaches. The notion of similarity is introduced, and different theories are elucidated. A generic framework is specified which describes the steps involved in measuring similarity between concepts in general. While this description is focused on the comparison of concepts expressed in description logics, it is adaptable to various other measures as well. To support this view, examples from several similarity theories are given.

A detailed review and comparison of similarity measures was given by Goldstone and Son [56], and Schwering [148]. A comparison between the SIM-DL similarity theory introduced within this thesis and related measures is given in chapter 5.

2.1 SEMANTIC SIMILARITY MEASURES

The theory of similarity has its origin in psychology and was established to determine why and how entities are grouped into categories, and why some categories are comparable to each other while others are not [56, 123]. The main challenge in *semantic* similarity measurement is the comparison of meanings as opposed to a purely structural comparison. A language has to be specified to express the nature of entities and a measurement theory needs to be established to determine how similar compared entities are. While entities can be defined in terms of attributes, the representation of concepts is more complex. Depending on the (computational) characteristics of the representation language, concepts are specified as (unstructured) bags of features, regions in a multidimensional space, or formal restrictions specified on sets using various kinds of description logics. While some representation languages have an underlying formal semantics (e.g., model theory), the grounding of several representation languages remains on the level of an informal description. As the computational concepts are representations of concepts in human minds, similarity depends on what is said (in terms of their formal representation) about these concepts. This again is connected to the chosen language, leading to the fact that most similarity theories cannot be compared. Besides the question of representation, context is another major challenge for similarity research. In many cases, meaningful notions of similarity cannot be determined without defining a context in which similarity is measured [44, 59, 91, 123].

Similarity has been widely applied within GIScience. Based on Tversky's feature model [168], Rodríguez and Egenhofer [143] developed the Matching Distance Similarity Measure that supports a basic context theory, automatically determined weights, and asymmetry. Raubal and Schwering [138, 149] used so-called conceptual spaces to implement models based on distance measures within geometric space (see also

[48]), while Sunna and Cruz [160] applied a network-based similarity measure for ontology alignment. Several measures [2, 3, 18, 32, 33, 79, 85] were developed to close the gap between (geo-)ontologies described by various kinds of description logics and similarity theories that had not been able to handle the expressivity of such languages. Other similarity theories [107, 129] have been established to determine the similarity between spatial scenes. The ConceptVISTA [46] (geo-)ontology management and visualization toolkit uses similarity for knowledge retrieval and organization. Klippel et al. [94] provided first insights into measuring similarity between geographic events.

A similarity theory comparing computational representations (of the concepts in our minds) in a cognitively plausible way consists of two layers. The representation layer specifies the language, i.e., a syntax and semantics, used to specify these concepts. The cognitive layer describes how concepts (and their representations, respectively) are compared. The cognitive layer answers questions such as which parts of concept descriptions are compared and how context (e.g., a certain task to be solved; see [82, 84, 85]) influences similarity judgments. While recent research from psychology and neurobiology argues for a situated nature of conceptualization and reasoning [10, 11, 12, 104, 121, 176], the concept representations used by most similarity theories from information science are stable, i.e., de-contextualized.

In the literature the following kinds of similarity measurement are distinguished: Feature-based, alignment-based, network-based, transformational, geometric, information theoretic, and model-driven. The following subsections focus on the first three kinds, as they are relevant for the presented SIM-DL similarity measure. It is important to note that most recent theories combine several approaches. For instance, the SIM-DL measure[1] is based on MDSM (which combines feature-based and network-based measures), network measures, and uses methods developed for alignment theories (details are given in chapter 5).

2.1.1 *Feature-Based Similarity*

The family of feature-based models, also called classic models, is the most prominent type of similarity measures. Most feature-based approaches use the ratio and contrast models introduced by Tversky [168]. While implementations and weightings differ, the underlying idea is that similarity can be expressed as a function of common and distinguishing features. Note that the term *feature* is used here as descriptor, i.e., aspect of a concept, not as representation of a real world entity (as in GIScience). Because of its major importance for GIScience and this thesis, the Matching Distance Similarity Measure (MDSM) is introduced here as an example for a feature-based model.

MDSM is the asymmetric and context sensitive semantic inter-concept similarity measurement approach developed by Rodríguez and Egenhofer [143]. It can be regarded as an extension of the model proposed by Tversky [168], and is therefore classified as a feature-based approach to similarity. MDSM distinguishes between three kinds of features: parts which are structural components of the individuals of a

1 Which itself is a model-driven measure.

class such as the riverbed of a river, functions which describe "what is done to or with a class" [143, p. 232] such as the function *transportation* is offered by canals (the idea of functions in MDSM is close to Gibson's [50] affordances), and attributes which are additional characteristics that can not be regarded as parts or functions such as the name of a canal.

$$S(c_1, c_2) = \omega_p * S_p(c_1, c_2) + \omega_f * S_f(c_1, c_2) + \omega_a * S_a(c_1, c_2) \tag{2.1}$$

Equation 2.1 shows the overall semantic similarity measure, which is regarded as the sum of the weighted similarities of the three kinds of features (parts, functions, and attributes) of the compared entity classes c_1 and c_2.

$$S_t(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha(c_1, c_2) * |C_1 \backslash C_2| + (1 - \alpha(c_1, c_2)) * |C_2 \backslash C_1|} \tag{2.2}$$

Equation 2.2 describes the asymmetric similarity function for each of the feature types. $S_t(c_1, c_2)$ is defined as the similarity for the feature type t between the entity classes c_1 and c_2. C_1 and C_2, respectively, are the sets of features of type t for c_1 and c_2, while $|C_1 \cap C_2|$ is the cardinality of the set intersection and $|C_1 \backslash C_2|$ is the cardinality of the set difference.

The relative importance α (equation 2.3) of the different features of type t is defined in terms of the distance d between c_1 and c_2 within a hierarchy that takes taxonomic and partonomic relations into account. *Lub* denotes the least upper bound, i.e., the immediate common superclass of c_1 and c_2 [143]. The distance is defined as $d(c_1, c_2) = d(c_1, lub) + d(c_2, lub)$.

$$\alpha(c_1, c_2) = \begin{cases} \frac{d(c_1, lub)}{d(c_1, c_2)}, d(c_1, lub) \leq d(c_2, lub) \\ 1 - \frac{d(c_1, lub)}{d(c_1, c_2)}, d(c_1, lub) > d(c_2, lub) \end{cases} \tag{2.3}$$

MDSM takes context into account. The weighting in the overall similarity function (equation 2.1) is calculated depending on the domain of application (discourse) using variability or commonality within the features (of each type). The context (C) is defined as a set of tuples over operations (op_i) associated with their respective nouns (e_j); see equation 2.4. These nouns correspond to MDSM entity classes, while the operations correspond to verbs associated with the functions defined for these classes [143]. A contextual specification such as $C =< (play,) >$, for example, expresses a domain of application that contains all classes which share the functional feature *play*.

$$C = \langle (op_i, \{e_1, ..., e_m\}), ..., (op_n, \{e_1, ..., e_l\}) \rangle \tag{2.4}$$

Within such a context the relevance (ω_t in equation 2.1) of each feature type is defined either by the variability P_t^v (equation 2.5) or commonality P_t^c function (equation 2.6) and then normalized with respect to the remaining feature types so that the sum of $\omega_p + \omega_f + \omega_a$ is always 1.

$$P_t^v = 1 - \sum_{i=1}^{l} \frac{o_i}{n * l} \tag{2.5}$$

The variability describes how diagnostic [57, 168] a feature type t is within a certain domain of application by assuming that the more characteristic each feature is for a given class the more diagnostic it is. A certain feature of type t has low relevance if it appears in many classes and high relevance if it is not common to the classes within the domain. P_t^v is the sum of the diagnosticity of all features of the type t in the domain and therefore 0 when all features are shared by all entity classes (P_t^v=1-1=0) and close to 1 if each feature is unique (where o_i is the number of occurrences of the feature within the domain) and the number of features l and classes n in the domain is high.

$$P_t^c = \sum_{i=1}^{l} \frac{o_i}{n * l} = 1 - P_t^v \tag{2.6}$$

Commonality is defined as the opposite of variability (Ptc =1- Ptv) and assumes that by defining a domain of application the user implicitly states what features are relevant [143].

In 2005, the author [77] extended MDSM to support role-filler pairs and thematic roles [155] to prevent wrong function matches. This extension also allows for partial matches within S_f.

2.1.2 *Alignment-Based Similarity*

Alignment-based approaches were developed as a reaction to shortcomings of feature-based (and geometric) models in representing structures. Both do not establish relations between features and dimensions, respectively. This also involves relations to other concepts or their instances. Using both models it is not possible to state that two concepts are similar, because their instances overlap with instances of another concept. As depicted in figure 1, the topological relation *above(circle, triangle)* [56] does not describe the same fact as *above(triangle, circle)*. During a similarity assessment participants may judge *above(circle, triangle)* to be more similar to *above(circle, rectangle)* than to *above(triangle, circle)* because of the same role, namely being above something else, the circle plays within the first examples (see also [119]).

The motivation behind alignment-based models is that relations between concepts and their instances are of fundamental importance to determine similarity [51, 53, 55, 120]. If instances of two compared concepts (per definition) share the same color, but the colored parts are not related to each other, then the common feature of having the same color does not influence similarity assessments. This means that subjects tend to focus on structures and relations more than on disconnected features. Hence, alignment-based models claim that similarity cannot be reduced to matching features, but determining how these features correspond to (*align with*) others [56].

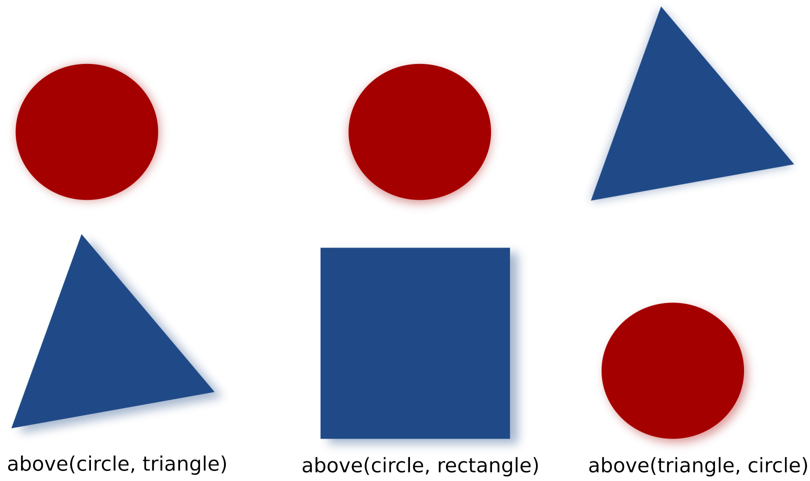

Figure 1: Being above something else as common feature used for similarity reasoning (see [56]).

From a set of available concept descriptors humans tend to select those for comparison which correspond in a meaningful way [40, 49, 53, 117, 120, 123]. The following three situations can be distinguished (see figure 3 for an example): alignable commonalities, alignable differences, and non-alignable differences. In the first case entities and relations match. For instance, in *above(circle,triangle), above(circle,triangle), above(circle,rectangle), and smaller(circle,triangle)*, the first two assertions are alignable because both specify an above relation, and common because of the related entities. In contrast, the second and third assertion form an alignable difference. While the assertions can be compared for similarity, the related entities do not match (but could still be similar). Non-alignable differences cannot be compared for similarity in a meaningful way. For instance, no meaningful notion of similarity can be established between *above* and *smaller*.

The Similarity as Interactive Activation and Mapping model (SIAM) is one possible alignment model (see also [88, 109]). It compares two scenes based on feature-to-feature and entity-to-entity node comparison. According to SIAM, the similarity between two scenes can be computed as described in equation 2.7

$$sim = \frac{\sum_{i=1}^{n} match\,value * A_i}{\sum_{i=1}^{n} A_i} \tag{2.7}$$

The symbol (n) stands for the number of nodes in the computation, while A_i is the activation of node i. The *match value* represents the similarity between the two features placed in correspondence according to the node i [56]. For a detailed description of SIAM and the underlying algorithm see [51, 53]. More details on structure mapping, alignment, and their computational characteristics were presented by Markman and Gentner [119]. Evidence for the suitability of alignment-based similarity for category-based induction (in contrast to feature-based models) was given by Lassaline [105].

2.1.3 *Network-Based Similarity*

Network-based models use semantic networks to represent entities, concepts, and relations as graphs (see figure 2 and 19). The best known network-based similarity measure is the DISTANCE measure introduced by Rada et al. [137]. Nowadays, these measures are mostly used as parts of more complex measures or toolkits such as MDSM, SIM-DL, or SimPack [17].

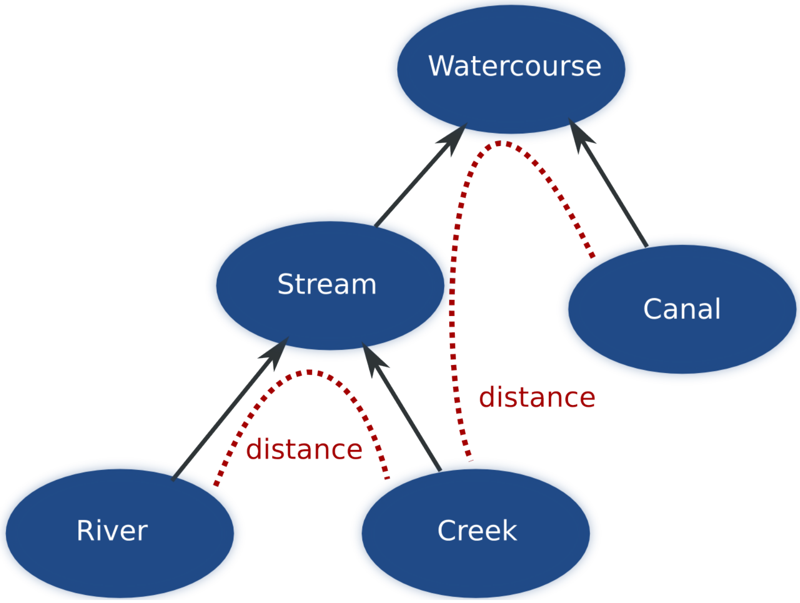

Figure 2: Conceptual distance in a subsumption hierarchy.

The hsw network measure [78] is introduced here, as it is relevant for this thesis (and based on Rada's DISTANCE measure). In contrast to theories assuming a constant distance within subsumption hierarchies, hsw proposes a variable weighting depending on the hierarchy depth. This reflects the fact that abstract concepts are less similar than concepts situated at a deeper level of the hierarchy. Such concepts already share all features of their ancestors. For instance, rivers and creeks differ with respect to features such as size, but share all features defined for streams and watercourses (see figure 2).

$$hsw(c_1, c_2) = \frac{depth(lub(c_1, c_2))}{depth(lub(c_1, c_2)) + edge\_distance(c_1, c_2)} \tag{2.8}$$

Equation 2.8 defines similarity as the ratio of the hierarchical depth level of the least upper bound (*lub*) of the compared concepts (c_1 and c_2) and the sum of this depth and the edge distance between these concepts. The edge distance is the shortest path; it is the number of edges which have to be passed from c_1 ands c_2.

2.1.4 *Similarity in Context*

The role of context in similarity measurement was examined by Rodríguez and Egenhofer [143] for feature-based measures and by Janowicz and Keßler [79, 85, 91, 92] for similarity measures based on various description logics. Turhan et al. [167] introduced a framework for processing context information based on modeling context as concepts using the Web Ontology Language OWL-DL. While this approach focuses on modeling the application (domain), other approaches (such as [25, 172]) propose a generic (top-level) context ontology. Jurisica [89] proposed a context-based similarity theory for information retrieval using the Similarity Query Language (SimQL).

The chapters 5 and 8 introduce specific contexts for SIM-DL and similarity measures in general; for now, we define context as any kind of information which does not describe the compared concepts directly but has an impact on similarity judgments during execution time.

2.2 SIMILARITY FRAMEWORK

By studying several similarity theories from information science and their application areas, we found generic patterns which jointly produce our framework for measuring similarity between concepts (see also [79, 81, 85]). The framework consists of the following five steps. Their concrete realization depends on the semantic similarity theory and the underlying representation language. Consequently, while some of these steps are important for a particular theory they may play a marginal role for another.

1. Selection of search (query) and target concepts
2. Transformation of concepts to canonical form
3. Definition of an alignment matrix for concept descriptors[2]
4. Application of constructor specific similarity functions
5. Determination of normalized overall similarity

As argued by Goodman [59], there is no global and application independent law on how similarity is measured. Every similarity theory should define in which way it implements the proposed steps and thereby specifies the semantics of similarity (values)[123].

2.2.1 *Context, Search and Target Concepts*

Before similarity is measured the compared concepts from the examined ontology (ontologies) have to be selected. Depending on the application scenario and theory, the search concept C_s can be part of the ontology or phrased using a shared vocabulary (in the latter case the term query concept C_q may be more appropriate) [79, 85, 111]. The target concepts $\{C_t\}$ are selected by hand or by the context of the query.

2 The term concept *descriptor* is used here as placeholder for feature, dimension, superconcept etc., which are used to describe the nature of a particular concept.

The context determines the domain of application [143] either by explicitly selecting the compared-to concepts or implicitly by defining a context concept C_c. In the latter case, the target concepts are all concepts subsumed by C_c[3].

The following list shows some exemplary similarity queries:

- How similar is *Canal* (C_s) to *River* (C_t)?
- What is most similar to *Waterbody* $\wedge$ *Manmade* (C_q)?
- Which kind of *Waterbody* (C_c) is most similar to *Canal* (C_s)?
- What is more similar to *Canal* (C_s), *River* (C_t) or *Lake* (C_t)?

In the first query, both concepts are part of an examined ontology. The target concept (*River*) is selected by hand. In the second query, the whole ontology is queried for concepts that are similar to a concept formed by the intersection of *Waterbody* and *Manmade*. The query concept is not necessarily part of the ontology itself but defined by the user. In contrast to the second query, the third query restricts the comparison to such target concepts that are subconcepts of *Waterbody*. Same as the query concept, the context concept is not necessarily a named concept within the ontology. The last query is an extended version of the first example with two target concepts (selected by hand).

In theory, one may also think of similarity measures without a given direction, i.e., without an explicit search and target concept, though this is hard to support from a cognitive point of view.

2.2.2 *Canonical Normal Form*

Before similarity is measured, the concepts have to be rephrased to a canonical normal form to eliminate unintended syntactic influence. Similarity should depend on what is said about concepts, not how it is said. If parts of compared concept descriptions (specified in a given language) denote the same facts using different language elements or statements, they have to be rewritten in a common form. This step mostly depends on the underlying representation language and its importance increases with the expressivity of the used language.

A simplified example can be derived from De Morgan's laws and propositional logic. The two expressions $\neg(p \wedge q)$ and $(\neg p \vee \neg q)$ are equivalent; before similarity is measured the following rewriting rule has to be applied.

Rewriting Rule 2.1
Condition: A concept description contains the expression $\neg(p \wedge q)$.
Action: Rewrite $\neg(p \wedge q)$ *to* $(\neg p \vee \neg q)$.

One may also think of canonization for other representations such as conceptual spaces. For instance, if the dimensions area, height, and width are part of a knowledge base. The category of things occupying $1m^2$, can be either expressed as a point on the area dimension or as a curve on the dimensions height and width. Per definition the denoted

3 For other kinds of contexts see section 8.4 and [14, 69, 91, 146].

category contains the same entities, but the similarity value would be 0 (using classical geometric-based similarity measures). In such a case, a rewriting rule has to map one representation to the other. Of course this example requires that the semantics of the involved dimensions is known.

2.2.3 *Alignment Matrix*

While the first step of the framework selects concepts for comparison, the alignment matrix specifies which and how concept descriptors (e.g., dimensions, features, super/subconcepts) are compared. The term alignment is chosen here, following research from psychology that investigates how structure and correspondence influences similarity judgments [40, 49, 53, 117, 120, 123]; see section 2.1.2. The term matrix points out that the selection of comparable tuples of descriptors requires a matrix $C_s^D \times C_t^D$ (where C_s^D and C_t^D are the sets of descriptors forming C_s and C_t, respectively).

Such an alignment matrix answers the following questions: In most similarity theories each concept descriptor from (C_s) is compared to exactly one descriptor from (C_t). How are these tuples selected? If the compared concepts are specified by a different amount of descriptors, how to treat surplus descriptors [139]? Does it make a difference whether the remaining descriptors are from the search or target concept? Are there special weights for certain tuples or are all tuples equally important? How similar are concepts to their superconcepts and vice versa? Does the similarity measure depend on the search direction; is it asymmetric [168]?

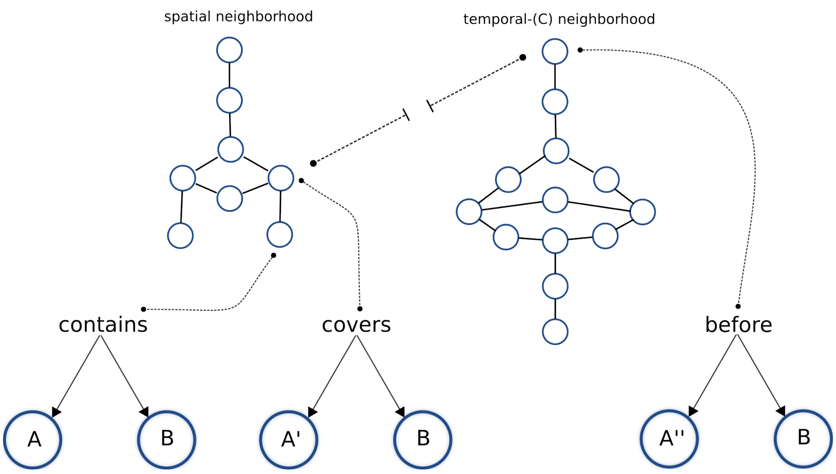

Figure 3: Topological and temporal neighborhoods within the alignment process.

For instance, taking topological and temporal neighborhood models into account, A and A' can be aligned while this is not possible for A and A'' in figure 3. The first tuple results in an *alignable difference*, while the second is a *non-alignable difference*. A meaningful notion of similarity can be derived in the first, but not in the second case.

In section 2.2.1, the distinction between search and target concept was introduced, and the kinds of queries that can be run against a similarity server were discussed. Next, the question how the search direction influences similarity has to be clarified. One can imagine the following scenarios (see also [147, p.111]). The user is searching for a concept that exactly matches the search concept (C_s) —

- ... every divergence reduces similarity.
- ... or is more specific.
- ... or is more general.
- ... or at least overlaps with C_s.

In the first case, similarity is 1 if $C_s = C_t$ and decreases with every descriptor from C_s or C_t that is not part of both specifications. Similarity reaches 0 if the compared concepts have no common descriptor. Asymmetry is not necessary, but can be achieved by weighting distinct features differently in dependence of whether they are descriptors of C_s or C_t. In the second scenario, similarity is 1 if $C_s = C_t$ or if C_t is a special type of C_s; else, similarity is 0. Such notion of similarity requires asymmetry. If C_t is a subconcept of C_s, the similarity $sim(C_s, C_t)$ is 1, while $sim(C_t, C_s) = 0$. The third case works the other way around, similarity is 1 if $C_s = C_t$ or if C_s is a special type of C_t. In the last scenario, similarity is always 1, except for the case that C_s and C_t do not even share a single descriptor.

In contrast to the first scenario, the remaining cases can be reduced to subsumption reasoning-based information retrieval as described by Lutz and Klien [111]. These scenarios only distinguish values between 1 and 0. In the second and third case, the search (query) concept is injected into the examined ontology. After reclassification, all subconcepts (respectively superconcepts) of C_s are part of the result set; see also figure 8. The last scenario can be solved accordingly by searching for a common superconcept of C_s and C_t.

Consequently, a similarity theory should be based on the first case or a combination of the first and second, or first and third case, respectively. Such combinations necessarily lead to asymmetric similarity measures. It is important to keep in mind that these design decisions are driven by the application area and not by a generic law of similarity [59, 60, 134, 154].

2.2.4 *Similarity Functions*

After choosing the compared-to concepts and aligning their descriptors, similarity is measured for each selected tuple $sim(X_{s_n}, Y_{t_m})$. Depending on the constructors used for X_{s_n} and Y_{t_m} different similarity functions have to be applied. The concrete functions depend on the similarity theory and the used representation language. For instance, in case of the Matching Distance Similarity Measure (see section 2.1.1), features are distinguished into different types during the alignment process (parts, attributes, and functions), however the same similarity measure can be applied to all of them. In most theories, each similarity function takes care of normalization (to values between 0 and 1) itself.

Assume the triples $role(X, Y)$ depicted in figure 3 are descriptors of the concepts A, A', and A''. To compare A and A', the similarity for the tuple $(contains.B, covers.B)$ has to be determined. This involves one measure for roles (arranged within a conceptual neighborhood) and one for their fillers. The similarity for such fillers is again derived from comparing their descriptors, i.e., role-filler pairs.

Note that, while we are focusing on inter-concept similarity here, certain similarity functions can also take knowledge about instances into account to derive information about concept similarity [32, 33].

2.2.5 *Overall Similarity*

In the last step, the single similarity values derived from applying the similarity functions to all selected tuples (of the compared concepts) are combined to an overall similarity. In most theories this step is a normalized (to values between 0 and 1) and weighted summation function.

For MDSM, the overall similarity is the weighted sum of the similarities determined between functions, parts, and attributes. The weights indicate the relative importance of each feature type using either a commonality or variability model (see section 2.1.1). At the same time, the weights act as normalization factors ($\sum \omega = 1$)[143]. For conceptual space-based approaches, the overall similarity is given by the normalized, i.e., z-transformed sum of compared values [138].

Finally, based on the previous steps, every similarity theory should state whether the similarity measure is reflexive, symmetric, transitive, strict, minimal, etc. (see [4, 30, 56, 134] for a detailed discussion from the perspective of computer science and psychology.). It is interesting to note that while mathematicians stress these properties to achieve a sound measure, researchers from cognitive science (especially psychology) and also artificial intelligence claim that most of these properties contradict with the nature of (human[4]) similarity judgments.

2.2.6 *Summary*

Summing up, the framework raises the following questions. Which concepts are compared? Is the comparison process directed? Is the similarity measure context-aware and asymmetric? How are the descriptors of compared concepts selected for measurement? How is the similarity between these descriptors determined? What is compared (measured)? Is the similarity normalized to values between 0 and 1? Are there different weights for particular descriptors? In answering these questions, a similarity theory defines its semantics, i.e., the interpretation of the values resulting from comparison.

4 Op de Beeck and colleagues also examined the role of asymmetry in stimuli comparison for monkeys (and other animals) [133].

3 ONTOLOGY

This chapter introduces ontologies and description logics as a representation language for ontologies.

3.1 ONTOLOGY -- MAKING SEMANTICS EXPLICIT

According to Gruber [61], "an ontology is an *explicit specification* of a *conceptualization*" used to achieve a *shared* and *common understanding* of a particular *domain* of interest (see also [63, 155, 159] and especially [64] on the principles of ontology). In the following paragraphs, we investigate each part of this definition.

EXPLICIT SPECIFICATION One of the most prominent distinctions between an ontology and other kinds of specifications (such as plain text definitions provided by a dictionary) is its explicit and formal character. An ontology is phrased using a formal language $\mathcal{L}$ which introduces the allowed symbols (the alphabet), the syntax (the grammar of $\mathcal{L}$), and the semantics by providing an interpretation function for *terms* phrased in $\mathcal{L}$. One has to distinguish between the semantics of pre-given symbols in $\mathcal{L}$ and the semantics of the terms phrased using the language $\mathcal{L}$ (see [63] for a formal characterization). The terminology phrased using the language $\mathcal{L}$ has to be satisfiable, i.e., specifications have to be contradiction free with respect to the TBox (see section 3.2.2 for details). An exception are the so-called (hierarchical) micro-theories (as used in OpenCyc), which allow to encapsulate inconsistent knowledge within an ontology. For instance, this allows to state that persons are living human beings and still describe former presidents of the US as persons.

Consider the following example for clarification: A dictionary describes a waterway as *natural or man-made linear body of water used for transportation*. From an ontological point of view this raises the following questions and restrictions [83]. In a formal language, *or* would be a logical (pre-given) symbol — a so-called constructor and could represent an exclusive or inclusive *or*, i.e., something is A or B, or both. Human readers would assume an exclusive *or* in this case, but an inclusive *or* for a definition such as *a canal is connected to natural or manmade waterbodies*. Next, if a waterway is used for transportation, can it be also used for different purposes in addition? This roughly corresponds to the existential restriction and value quantification constructors in description logics (see section 3.2). Is it sufficient to be built to potentially transport something or does a particular waterway has to transport at least one individual right now (at all times)? Finally, while this plain text definition looks plausible at first glance, it puts some problematic restrictions on waterways. As a waterway is a linear body of water, rivers can be waterways, while most lakes cannot. A more detailed example from the domain of hydrology, highlighting the described problems, is provided in chapter 4 (and 7).

CONCEPTUALIZATION Not every set of terms specified in a formal language is an ontology. Ontologies are about concepts. This sounds trivial at first, but is one of the main characteristics of ontologies. Depending on the used language and framework, an ontology is made of the following components: classes (types), relations, functions, attributes, rules, axioms, and instances. For this thesis, classes and relations (binary predicates in case of description logics) are of major importance. If we state that ontologies are about concepts, we refer to the mental handles used by humans for communication and reasoning (see also chapter 4). These concepts are the link between the symbols in our language and the (real world) referents (entities). Naming classes, i.e., explicit specifications within a TBox, establishes the same kind of link as for the mental handles. In the same way as concepts refer to real world entities, classes refer to instances (individuals). For this reason, the relation between concepts and classes if often depicted using the double semantic triangle (see figure 4) initially introduced by Sowa [155] (based on the work from Ogden and Richards [132]).

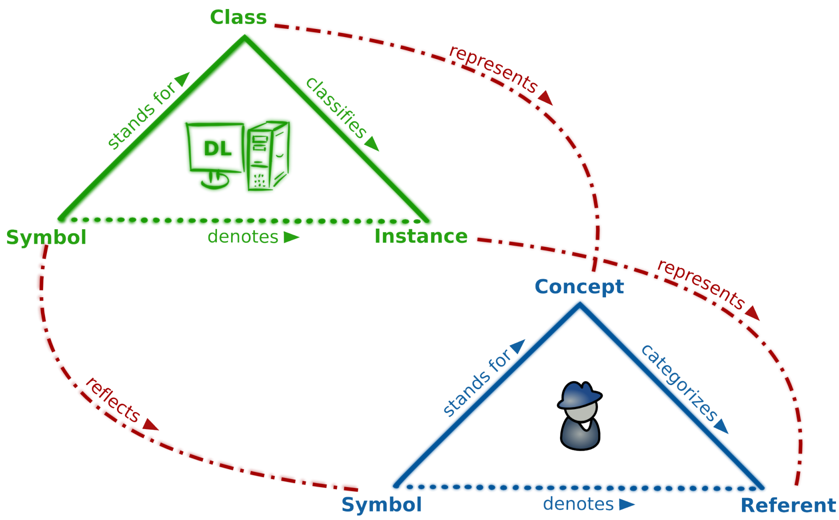

Figure 4: The double meaning triangle (based on Sowa [155]).

The triangle of reference describes the relationship between symbols, concepts, and referents. Instead of a direct link from symbols to referents, symbols stand for concepts (in our minds). These concepts are used to categorize referents which are alike[1].To make this relation clear, the terms *category* and *concept* are distinguished in the literature [115]. A category is the set of entities, while a concept is the schema applied to classify entities into categories. The dictionary definition of waterway presented above is a textual description of the concept *Waterway*, while all real world entities categorized as waterways form the category of waterways.

The question emerges how the human and machine-based triangles are related. Instead of trying to merge at least one of the corners of both

1 This is the point where similarity comes into play.

triangles, we leave them separated and introduce relations between them. This is for the following reasons: Concepts and classes cannot be merged, as the classes are models of the concepts in our mind. Classes represent concepts using a formal language. One could also argue that classes approximate concepts, which also points to quantitative aspects of this relationship (leading to the question how good the approximation is; see figure 6 for more details). The same arguments hold between real world entities and knowledge base instances. This raises the problem that the extension of the concept *Waterway* is different from the extension of the class *Waterway*. Finally, this is the reason why the symbol corners should not be merged. Besides syntactical issues, symbols should not be merged to keep the difference between representation and represented in mind. Instead, the symbols used as names for the classes reflect or mirror the symbols used for the concepts.

Nevertheless the terms concept and entity are used for the computational representations within the description logics and ontology engineering community. For reasons of readability and following this convention, we will use the terms concept and class, as well as entity and instance as synonyms. If necessary (as in chapter 7) we will clearly point out whether we refer to ontological concepts or those in human minds.

shared and common understanding Ontologies are developed for a specific purpose which can be used to distinguish between several types of ontologies: Ontologies can be classified by their degree of generalization (abstraction) and by the domain of interest they cover. Mostly, the distinction between top-level, domain, task, and application ontologies (see figure 5) is made [62, 63]. This classification takes the granularity and the thematic scope of the ontology into account. Top-level ontologies, such as the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [122], cover knowledge about the world in an application and domain independent way. They focus on the distinction between abstract and physical entities or endurants and perdurants. Together with additional reasoning services, top-level ontologies may be the foundation of semantic reference systems as described by Kuhn [100]. Ontologies from lower levels should refer to these terms to improve interoperability.

Domain ontologies deal with concepts that are used in a particular domain, e.g., hydrology. The geographic feature type ontology which is introduced in chapter 4 could be considered as domain level ontology. It provides the concepts necessary to align [131] feature types from local applications (such as gazetteers or Web services). Domain ontologies form the intermediate level between the abstract concepts specified within top-level ontologies and the concrete implementation-centered view of so-called application ontologies. Task ontologies describe concepts that are needed to fulfill a special task; this may be an ontology specifying the concepts involved in a particular function offered by an information system. However, and in contrast to the other types, this term is uncommon. Additionally, the boundaries between these types of ontologies are rather fuzzy.

The second common classification distinguishes local and global on-

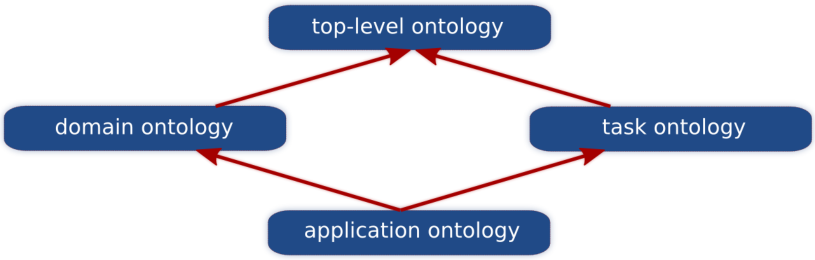

Figure 5: Four types of ontologies and their specialization relationships (from Guarino [62]).

tologies [169]. The assumption that global ontologies are top-level ontologies and local are domain-level ontologies is false. It is not possible to map these classifications directly, because they focus on different aspects. A global ontology can be on the domain level, but also on the level of a single company. For the characterization of global ontologies it is important that they are the common agreement of multiple units or working groups. If all departments of an international company agree on a shared conceptualization for their vocabulary this may be called a global ontology of this company. Local ontologies have to be aligned to other local ontologies or to a global ontology. In contrast, the global ontology does not require connections to other (global) ontologies, because it acts as reference [100]. Same as for the classification introduced before, the distinction between global and local ontologies is fuzzy.

More important than the distinction between local and global is the question of how these ontologies are used. Even if the idea of having a commonly agreed upon global (or top-level) ontology seems to be promising and worthwhile, it remains impractical until now. As pointed out by Uschold [169], every community and working group focuses on a special view of the world and wants to keep its own familiar vocabulary and definitions for good reasons. In fact, talking about heterogeneous structures, such as the semantic Web, we have to deal with a high amount of local ontologies. These ontologies, even when describing the same real world features, can neither be merged nor translated automatically in many cases. There is often no agreement or shared understanding between different groups or communities. Consequently, one should focus on techniques such as similarity, alignment, and context-dependency to build bridges between local communities (based on their local conceptualizations).

DOMAIN As depicted in figure 6, an ontology describes a certain domain specific world view by trying to restrict the set of possible models to the so-called *intended model*. As described by Guarino [63], in most cases the intended model is a subset of the ontology, i.e., the ontology captures more than it should. The ontology itself is a subset of the set of possible models for the given language and conceptualization.

This makes the selection of a particular representation a central requirement for reasoning [34]. Essentially, every class hides or highlights some aspects of a human's conceptualization. In choosing a rep-

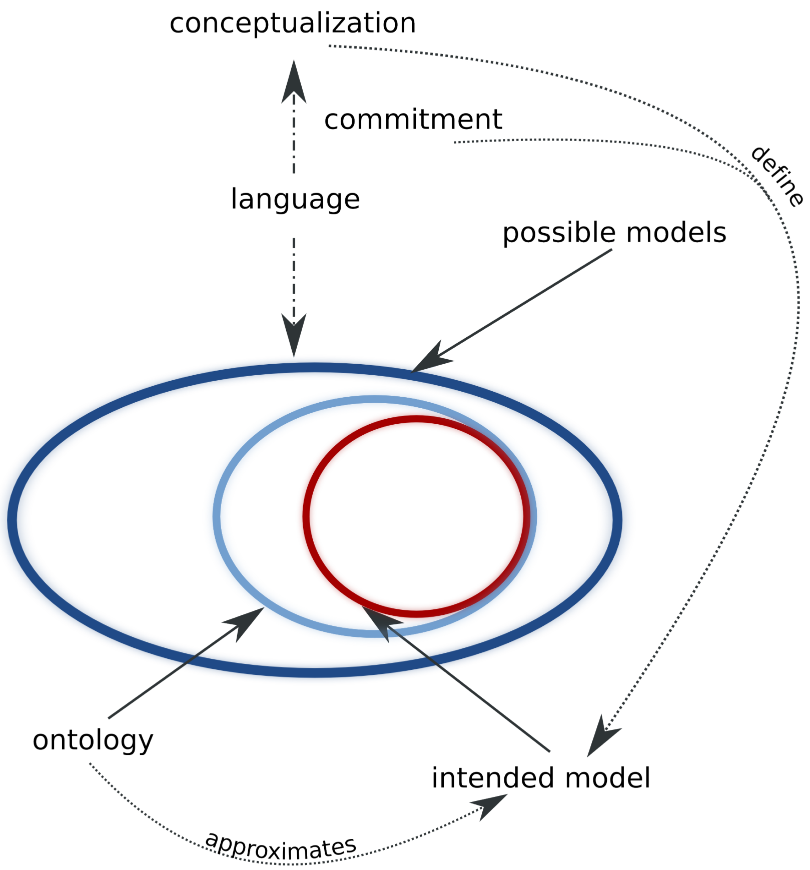

Figure 6: The conceptualization (and commitment) defines the intended model as subset of all possible models. By approximating the intended model, an ontology tries to capture (reflect) this commitment (based on Guarino [63]; see also [136]).

resentation one decides how to carve up the domain of discourse, i.e., to make some *ontological commitments* [63]. If, e.g., a reasoner returns that all lakes are linear, neither the ontology nor the reasoner is wrong, but the classes specified in the ontology do not approximate the intended model. Consequently, one has to find a better representation to match the domain specific needs.

3.2 Description Logics

Description Logics (DL) are a family of knowledge representation languages used to model concepts and entities in a knowledge base (e.g., an ontology). Such a knowledge base consists of a TBox ($\mathcal{T}$) containing the terminology, i.e., the vocabulary describing a given domain, and an ABox storing assertions (about named entities). Description logics distinguish two kinds of symbols, logical and non-logical. The former have a pre-defined meaning grounded in set theory, while the latter are domain specific. Logical symbols are either[2] constructors ($\sqcap, \sqcup, \exists, \forall, \leq, \geq$) used to compose complex concepts out of primitive ones, or $\equiv$ and $\sqsubseteq$ for equality axioms and inclusion axioms, respectively. Same as for first order logic (FOL), the formal semantics of description logics is given by their interpretation. An interpretation $\Im$ is defined as a tuple $\langle \Delta^{\mathcal{I}}, \mathcal{I} \rangle$. $\Delta^{\mathcal{I}}$ denotes a non-empty set called the domain of interpretation, whereas $\mathcal{I}$ describes the interpretation function from non-logical symbols to elements and (binary) relations over $\Delta^{\mathcal{I}}$. The subset $C^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ associated with a concept C is also called its extension. Within this thesis, the terms *description* and *specification* of a concept denote the statements (phrased using a particular description logic; see table 1) used to specify this concept.

The most famous application of description logics is within the Web Ontology Language standard (OWL) [16]. OWL comes in different flavors whose semantics is defined according to a particular DL: OWL-Lite is built on the description logic $\mathcal{SHIF}$, while OWL-DL corresponds to $\mathcal{SHOIN}(\mathcal{D})$. The extended new version OWL 1.1 matches the expressivity of $\mathcal{SROIQ}(\mathcal{D})$. OWL is the de facto standard for Web ontologies and supported by various reasoners and ontology editors.

3.2.1 *The $\mathcal{ALCHQ}$ Language*

For this thesis, the description logic $\mathcal{ALCHQ}$ is chosen, because it is close enough to OWL-DL, leaving aspects that are not relevant for similarity aside. $\mathcal{ALCHQ}$ even supports qualified number restrictions which are part of the new OWL 1.1. The main difference between $\mathcal{ALCHQ}$ and OWL-DL is the missing support for several role (i.e., binary predicate) axioms such as role inclusion (a similarity measure for role intersection has been discussed by Janowicz [79]), role transitivity, and inverse roles on the one hand, as well as nominals and datatype properties on the other hand. While it is difficult (and questionable) to find a meaningful notion of similarity for role axioms such as transitivity, the similarity between nominals (and simple datatypes) boils down to instance similarity.

As described in table 1, $\mathcal{ALCHQ}$ supports intersection, union, full existential quantification, value restriction, full negation, and qualified number restrictions. In this work, the letters A and B are used for atomic concepts, R and S for roles and C and D for complex (composed) concepts. X and Y are used for general statements about similarity and alignment that hold for both concepts and roles. Additional background information about $\mathcal{ALCHQ}$ and related description logics is discussed by Baader et al. [6]. Details about canonization and rewriting rules for $\mathcal{ALCHQ}$ are discussed in chapter 5.

3.2.2 *Reasoning*

Reasoning is the most important motivation for developing ontologies [73]. The most important TBox reasoning services are subsumption and satisfiability[3], but also matching, unification, and concept rewrit-

2 Leaving punctuation and numbers aside.
3 Note that subsumption can be reduced to (un)satisfiability.

Table 1: Syntax and semantics of $\mathcal{ALCHQ}$.

Syntax	Semantics	Name
$\top$	$\Delta^{\mathcal{I}}$	Top
$\bot$	$\emptyset$	Bottom
A	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$	Atomic concept
R	$R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$	Atomic role
$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$	(Full) negation
$C \equiv D$	$C^{\mathcal{I}} = D^{\mathcal{I}}$	Concept equality
$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$	Concept inclusion
$R \equiv S$	$R^{\mathcal{I}} = S^{\mathcal{I}}$	Role equality
$R \sqsubseteq S$	$R^{\mathcal{I}} \subseteq S^{\mathcal{I}}$	Role inclusion
$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$	Concept intersection
$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$	Concept union
$\forall R.C$	$\{a \in \Delta^{\mathcal{I}} \| \forall b.(a,b) \in R^{\mathcal{I}} \to y \in C^{\mathcal{I}}\}$	Value restriction
$\exists R.C$	$\{a \in \Delta^{\mathcal{I}} \| \exists b.(a,b) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$	Existential quantification
$\leq nR.C$	$\{a \in \Delta^{\mathcal{I}} \| \|\{b \in \Delta^{\mathcal{I}} \|(a,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}\}\| \leq n\}$	Qualified max. number restriction
$\geq nR.C$	$\{a \in \Delta^{\mathcal{I}} \| \|\{b \in \Delta^{\mathcal{I}} \|(a,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}\}\| \geq n\}$	Qualified min. number restriction

ing (see [128] for details). Typical ABox reasoning tasks involve the retrieval of all individuals of a given concept, instance checking, and consistency checking. Besides these classical reasoning services, non-standard inference plays a major role in recent reseach on description logics. This especially includes the least common subsumer and most specific concept [102], as well as approximation [20], and similarity reasoning [2, 3, 18, 32, 33, 79, 85].

Instead of discussing the theoretical background and possible implementations of each reasoning service, the definitions of subsumption and satisfiability are given here, because the SIM-DL similarity server implements a so-called ABox tableaux algorithm to compute subsumption between concepts based on ABox satisfiability (see chapter 6). Interested readers are referred to the overview given by Baader and Sattler [5], which was used as the basis for this implementation. Additional details, relevant for semantic similarity measurement, are discussed in chapter 5.

Definition 3.1 *Subsumption: The concept C is subsumed by the concept D with respect to the TBox $\mathcal{T}$, written as $C \sqsubseteq_{\mathcal{T}} D$, if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ for every model of $\mathcal{T}$.*

Definition 3.2 *Satisfiability: The concept C is satisfiable with respect to the TBox $\mathcal{T}$, if there exists a model of $\mathcal{T}$ such that $C^{\mathcal{I}} \neq \emptyset$.*

3.2.3 *DIG-Interface*

The DIG interface is an API for reasoning in DL-based systems [15]. The DIG 1.1 specification provides an interface for reasoning services based on the $\mathcal{SHOIN}(\mathcal{D})$ language. The specification provides an XML-encoded HTTP interface. Clients communicate with a server via HTTP `POST`, with requests and responses encoded based on the underlying DIG XML Schema[4]. DIG distinguishes between different types

4 The DIG XML Schema can be found at: http://dl-web.man.ac.uk/dig/2003/02/.

of messages and operations. The reasoner's *identification* message is comparable to OGC's `getCapabilities` request: the server responds with the language and services it supports. This is especially important because of the variety of DL languages, i.e., not every DIG server will support all constructs that are part of the specification (the basic constructs are compulsory, however). The *management* operation creates or releases a knowledge base (KB) that is further identified with a unique URI. *Tells* operations insert assertions into the reasoner's KB, while *Asks* operations allow the client to perform reasoning tasks on the KB (see [15] for details).

Most ontology editors, frameworks, and reasoners implement the DIG interface, which results in a loosely coupled client-server infrastructure for semantic Web applications. Consequently, the reasoner does not need to run on the same machine as the application, or ontology. An outlook on a distributed service infrastructure based on DIG is given in section 4.2. Several extensions are required to enable DIG-based communication with similarity servers. These extensions are discussed in section 6.2.

4 USE CASE

This chapter introduces the gazetteer application area employed within this work to describe and evaluate the similarity theory. The use case is also used to demonstrate how similarity can be applied for information retrieval and to improve the accessibility of complex systems (e.g., ontologies, knowledge organization systems) for human users. The chapter starts by introducing gazetteer research. Next, the vision of a distributed gazetteer infrastructure is elucidated and the role of similarity is discussed. A geographic feature type ontology is introduced to demonstrate how to map between geographic feature type thesauri used within gazetteers and the ontologies necessary for similarity reasoning. A new similarity-based gazetteer Web interface is presented.

4.1 GAZETTEERS IN GISCIENCE

Gazetteers are place name directories containing names, spatial references, feature types, and additional information for named geographic places. They are key components of all georeferenced information systems, including GIScience applications in many diverse fields, Web-based mapping services, and the emerging Web 2.0. A typical use case for gazetteers is information retrieval where queries can be based on place names or coordinates. They are central to the process of geoparsing where references to geographic locations by place name are recognized in texts and converted to coordinate references (see [70]). Gazetteers are also components of complex reasoning services such as the identity assumption service for historical places discussed by Janowicz [78] (see section 8.5.1). From an information theoretic point of view, a gazetteer record can be defined as a triple (N, F, T) where N corresponds to one or more place names, F represents one or more geographic footprints (i.e., locations), and T is the type of the described feature (i.e., place). In the context of gazetteers, a feature is a (representation of a) real world entity. The feature type which is selected from a typing scheme or ontology (as a concept) is used for feature categorization. A named geographic place is an abstract entity defined to refer to a physical region (extent) in space and categorized (typed) according to commonly agreed characteristics. *Place* is a social concept of interest for a particular community during a certain time span. Its name is a symbol used for communication.

Categorization is a central cognitive process. This section focuses on two reasons for the categorization of places: communication and cognition. Categorizing into types improves communication about places with which at least one communication partner is unfamiliar. For instance, when giving directions such as: "follow the *path* along the *river* up to the *bridge*, then turn right towards the *market place*." Typing is also the key to prediction, reasoning, and decision making which all require an abstraction from entity to type level. What humans experience as a place is, in fact, the set of perceivable characteristics of the region in space the place refers to by its type and name (see also [24]). This includes the surface and texture of the physical region of earth, man-made entities such as buildings, and knowledge obtained from maps, books, and other information sources. Beyond those perceivable characteristics, places may also be typed by convention, such as administrative areas. The referenced region (or entity) can also be described in relation to other regions or entities, such as "East Frisia is a coastal region in the northwest of Lower Saxony". The definition of place as a mental handle pointing to real world regions (or entities) is independent of a specific name or an affixed and stable portion of space.

Since the name acts as a symbol for communication, a particular place can be referred to using various names by different people and in different ways through time. This also includes placeholders such as "Anyshire". The spatial extent referred to by the place name may vary over time or be known only in a general sense. The clear distinction between real world and reference also helps to explain how places can disappear without causing inconsistency. One can argue that a place no longer exists when there is no human left who is aware of this place. A place, such as a temporal Normand settlement, moves when the perceivable characteristics move (as opposed to the region on the earth's surface)[1].

The partition into names, footprints, and types corresponds to the minimum definition of a gazetteer entry. The full set of descriptive elements also includes details such as spatio-temporal history[2]. The name of a place is called its textual reference, while the footprint is called the spatial reference. Thus, a gazetteer supports at least two functions. First, it maps between place names and respective footprints: $N \longrightarrow F$; and second, between names and types: $N \longrightarrow T$. Several online gazetteer services support queries by place name, footprint, and type via a Web page or through an application programming interface (API). These functionalities are integrated into other online services; for example, to translate a place-name query into a footprint query in order to search data sets where only spatial access is supported.

4.2 TOWARDS A DISTRIBUTED GAZETTEER INFRASTRUCTURE

The long term vision of current gazetteer research is the development of a distributed local-responsibility service infrastructure instead of a single world gazetteer [83]. Such an infrastructure can be compared to the Domain Name Service (DNS) which maps hostnames on the Internet to their IP addresses. Each gazetteer offers lookup for local places within its spatial and thematic scope. If the gazetteer cannot answer a request, it redirects the query to a higher level gazetteer which decides whether it or another gazetteer can resolve the query. The underlying idea is that gazetteers should contain and maintain data of interest for

1 This leads to the question of place identity which is out of scope for this work.
2 For instance, the ADL Gazetteer Content Standard allows for a *Time Period Note* for names, spatial footprints, and types.

the community running the service. This ensures that the stored data is both accurate and up-to-date.

A distributed gazetteer infrastructure raises several challenges for both the georeferencing and the type-lookup function. For georeferencing, the main challenge is that several names may point to the same place using different footprints, which includes divergences between the referred-to coordinates, but especially between the kind of footprint such as point versus polygon representation (see [76]). In the case of type-lookup, one must ensure that all involved gazetteers share a common understanding of the feature types used. Gazetteers are developed for different thematic scopes and spatial scales. This may require different conceptualizations of the described features. A common feature type specification needs to be generic enough to form a top-level for all gazetteers and extensible to allow for local type definitions.

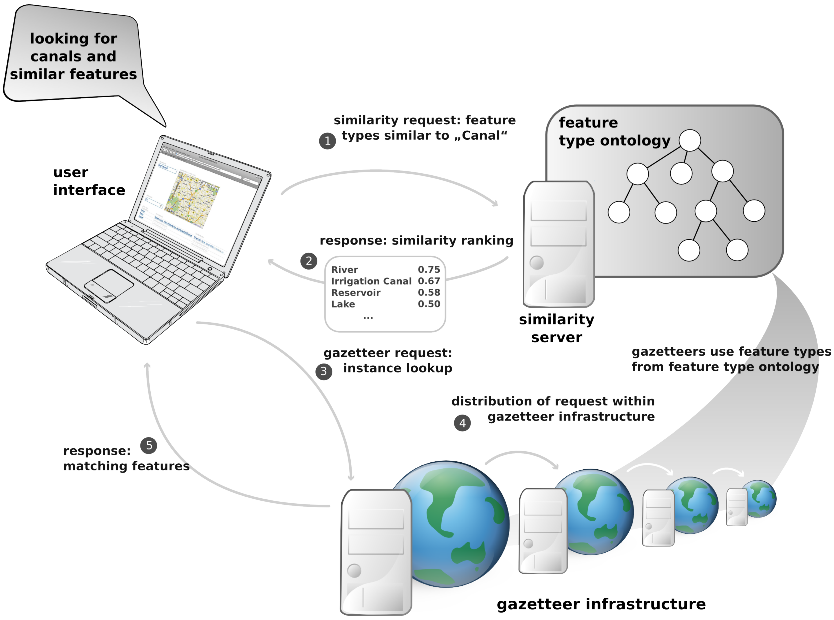

Figure 7: Similarity-based feature type lookup within the proposed gazetteer infrastructure (based on Janowicz et al. [85]).

Figure 7 illustrates the role of a similarity server (such as the SIM-DL server introduced in this thesis) within the proposed gazetteer infrastructure. In a first step, the user chooses the kind of feature type. At first glance, this does not differ from the original gazetteer type-lookup query. However, instead of querying gazetteers directly, the query is handled by the similarity server. The similarity server performs two kinds of queries, one for supertypes [111], i.e., superconcepts, of the searched geographic feature type, and one for similar types [85] (see figure 10). All types used by the local gazetteers are either types defined within the global [169] feature type ontology or added as subtypes by these local gazetteers. The server responds with a descending list of similar types, starting with the most similar type with respect to the user's query. For instance, if a user is searching for the type *Canal*, the server would also propose *River* and *Irrigation Canal* [85]. In the next step, the user selects the geographic feature type which should be used for the type-lookup query. Finally, the query is directed to the local gazetteers which deliver the appropriate features, e.g., rivers.

Figure 8 points out the difference between a subsumption and a similarity query within the gazetteer infrastructure. For reasons of simplification the similarity server (which acts as kind of a middleware) is not depicted. This corresponds to the case where similarity is computed on the user's machine.

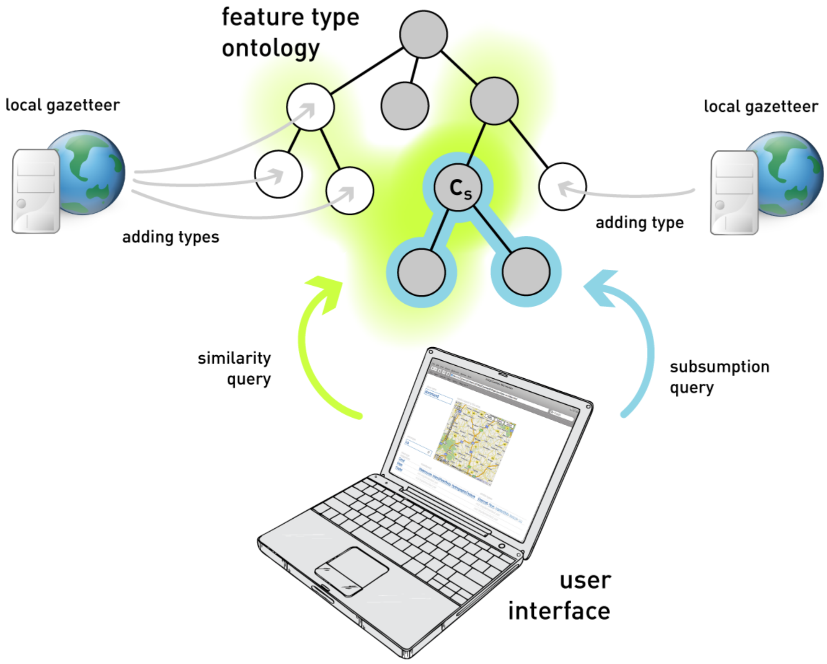

Figure 8: Subsumption and similarity-based information retrieval within the proposed gazetteer infrastructure (from Janowicz and Keßler [83]).

Subsumption-based reasoning has its origins in computer science and especially within knowledge representation. It is the most prominent of several inference techniques used within ontology-based information retrieval. The idea behind subsumption-based retrieval, as described by Lutz and Klien [111], is to rearrange a queried application ontology [63] taking a search concept (C_s) into account, and to return a new taxonomy in which all subconcepts of the specified search concept satisfy the user's requirements. This approach forces the user to ensure that the search concept is specified in a way that it is neither too generic (and therefore at a top level of the new hierarchy) nor too specific to get a sufficient result set. In fact, the search concept is not the searched concept (see [87]). It is a formal description of the minimum characteristics all retrieved concepts need to share. In most cases, users do not choose an existing concept from the ontology but create the search concept using primitives (e.g., a shared base vocabulary [111]). However, there are also situations where it could make sense to choose an existing concept as search concept.

For instance, a user searching for *Waterbody* could get *Canal*, *Ocean*, *River*, *Reservoir*, *Lake*, *Channel*, etc. as result set. The other way around, using *Canal* as search concept would only return *Canal* as result (if we assume that there is no subconcept of *Canal* in the geographic feature type ontology). The challenge for human computer interaction becomes apparent in the case of canals in California (USA). More than 40 of them are named channel, while they are typed as *canals* and not *channels* in the Alexandria Digital Gazetteer (ADL). According to the ADL Feature Type Thesaurus (FTT)[3], channels are natural while canals are manmade. A good example is the Lytle Creek Channel which is neither a creek nor a channel, but a canal (according to ADL).

In case of similarity reasoning, the concepts the users type into the system are the concepts they are looking for, i.e., the searched concepts. In contrast to subsumption, the benefits similarity offers during information retrieval, i.e., to deliver a flexible degree of conceptual overlap to the searched concept, stand against shortcomings during the usage of the retrieved information, namely that the results do not necessarily fit the users' requirements. To make the difference between both approaches more evident, one can imagine a search concept specified to retrieve all concepts whose instances *overlap* with waterways. In contrast to the subsumption-based approach, similarity measurement would additionally deliver concepts whose instances are located *inside* or *adjacent* to waterways, and indicate through a lower degree of similarity that these concepts are close to, but not identical with the users' intended concept.

In figure 8, the difference is depicted by crisp borders in the subsumption case and a cloud in case of similarity. It indicates that the result is a descending list of proximity values describing how close particular feature types are to the search concept.

4.3 TOWARDS A FEATURE TYPE ONTOLOGY

Georeferencing is the core functionality of gazetteers as place name directories. The distinction between different feature types is enabled by thesauri, which contain semi-formal descriptions of the feature types and can be queried using the type-lookup functionality. The expressivity offered by these thesauri is not sufficient for the proposed distributed gazetteer infrastructure. To fully support subsumption and similarity-based reasoning, a transformation of the thesauri into ontologies is required. This section highlights some of the steps and design decisions taken to convert a thesaurus into a feature type ontology. In the following, we use the ADL Feature Type Thesaurus as example; the procedure can be transferred to other thesauri. A more detailed discussions is given by Janowicz and Keßler [83].

Thesauri are defined as controlled vocabularies with a fixed number of relationships. These relationships specify hierarchies, associations, and equivalences. The hierarchical relationships can be further specified as being generic (is-a), partitive (whole-part), or instantiative (describing the relation between an instance and its type). Thesaurus stan-

3 http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/index.htm

dards [1] allow multiple hierarchies (i.e., a concept can occur in more than one hierarchical tree), but most thesauri (such as the ADL FTT) use a single is-a hierarchy to simplify maintenance and the display of relationships. This single inheritance structure forces every term to be a *narrower term* of only one of the six top terms of the ADL thesaurus; for example, *cities* are only classified as *administrative areas*, but not as *manmade features*. In addition, most thesauri do not use the hierarchical relations in such a strict sense as ontologies, i.e., they are not directly comparable to the subconcept and superconcept relations in ontologies [170]. Therefore, (is-a) transitivity cannot be taken for granted. An example is the term *hydrographic structures* defined in the ADL Feature Type Thesaurus as "constructed bodies of water". The subterm *canals* fits this definition, while the subterm *offshore platforms* does not. Therefore, searching for *hydrographic structures* using the ADL gazetteer Web interface[4] also returns *offshore platforms* (which are not bodies of water).

The associative *related term* relation is used to express diverse kinds of relations between terms, so that its semantics remain ambiguous – for instance, *related term* is used to describe the relation of *lakes* to *reservoirs*, i.e., a functional relation, but also to *wetlands*, i.e., a topological relation. The associative relationship can also point to proximity between terms which could not be described using the equivalency relationship. The associative relations in thesauri are not defined in any way that is transferable to ontologies (although the most recent thesaurus standards present common subtypes of associations). Instead, ontologies define the type of association explicitly, which allows for additional reasoning capabilities. Using the example above, an ontology would state that offshore platforms are *located within*[5] bodies of water.

Equivalency is used to introduce alternative terms that describe a term that is semantically equivalent within the scope of the particular thesaurus. This is the reason why thesauri are called controlled vocabularies. One term (the preferred term) is chosen to represent a concept while other possible terms (non-preferred terms) are entered as equivalent terms. These alternative terms are not part of the controlled vocabulary but are considered to be lead-in terms which lead to the appropriate controlled vocabulary term. For instance, *lagoons* are a non-preferred term leading to the preferred term *lakes*. Additionally, the ADL FTT contains textual definitions for *preferred* terms in the form of so called *scope notes*. The decision which term is preferred and which not is also driven by the question whether the differences between these terms can be made explicit. As the expressivity of ontologies is higher, it makes sense to also create concepts for non-preferred terms [83].

It has to be pointed out that the structure of the ADL FTT is not wrong or badly designed, since thesauri are developed for different purposes than ontologies. However, there is a lack of formalization and explicit semantics from an ontological point of view, which makes an automatic transformation into a feature type ontology impossible. To manually transform the thesaurus and preserve the original naming and structure, the syntactic and semantic conversion framework de-

4 http://www.alexandria.ucsb.edu/clients/gazetteer/

5 Which could be mapped to the topological *inside* relation; however, one has to keep in mind that this neglects the shift from 2D to 3D.

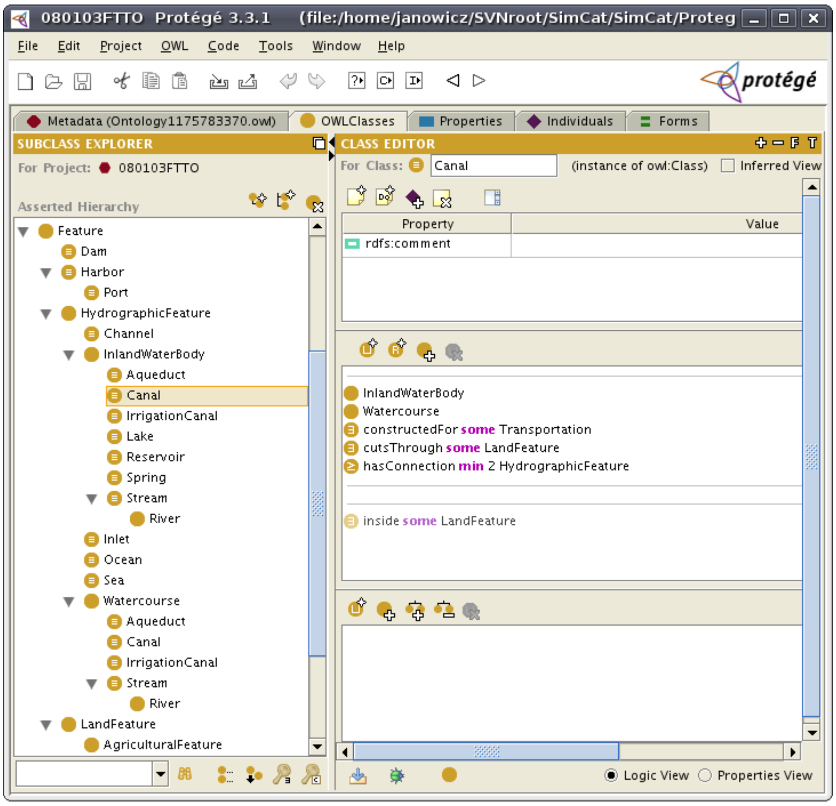

Figure 9: A partial view of the feature type ontology (as of January 3rd, 2008).

scribed by van Assem et al. [170] can be used. The resulting ontology[6] (see figure 9 for an extract) uses the top level concept *Feature*, subsumed by different classes such as *Manmade*, *HydrographicFeature*, or *Transportation*; note that these classes are not disjoint, i.e., the concept *Canal*, for example, is a subclass of *Manmade* and *HydrographicFeature* at the same time. Moreover, feature types can be related to each other with an arbitrary number of (DL) roles which have to be extracted manually from the *related term* relations and the scope notes in the thesaurus. For example, we introduce the role *hasConnection*, with the sub-roles *hasOrigin* and *hasDestination*, to specify that a canal connects at least two hydrographic features.

Another interesting aspect of mapping from thesauri to ontologies is partonomy. The ADL Gazetteer Content Standard (GCS) [71], on which the ADL gazetteer structure is based, allows the establishment of relationships between gazetteer entries. The existing ADL gazetteer has implemented only one relationship, the *part-of* relation between features (not types). This is an administrative part-of relationship, not a spatial one, although an administrative relationship infers spatial containment in many cases. From an ontological point of view, both part-of relations have to be distinguished [175] and defined within the

6 The ontology is under development, recent versions can be downloaded from the SimCat project page at: http://sim-dl.sourceforge.net/downloads/.

feature type ontology. This would allow to express that Münster is part-of Germany, but also that cities are part-of countries (or, in case of spatial containment, that islands are contained in bodies of water).

This discussion of the conversion process shows that the generation of a geographic feature type ontology requires significant effort. In the following, by introducing a new gazetteer Web interface based on similarity (and subsumption), we argue that this conversion is worthwhile.

4.4 SIMILARITY-BASED GAZETTEER WEB INTERFACE

To efficiently use the ADL gazetteer's Web interface, users currently need detailed knowledge of the FTT hierarchy to select the adequate preferred term for what they are looking for. If the users are not aware of the FTT hierarchy, retrieving the desired information is complicated and tedious, as the users have to consult the FTT first to find out about the preferred term for the query. Including the non-preferred terms used for navigation purposes, this adds up to more than 1200 terms with short, often ambiguous scope notes. To overcome these difficulties, we propose a subsumption and similarity-based gazetteer Web interface based on the introduced feature type ontology as shown in figure 10.

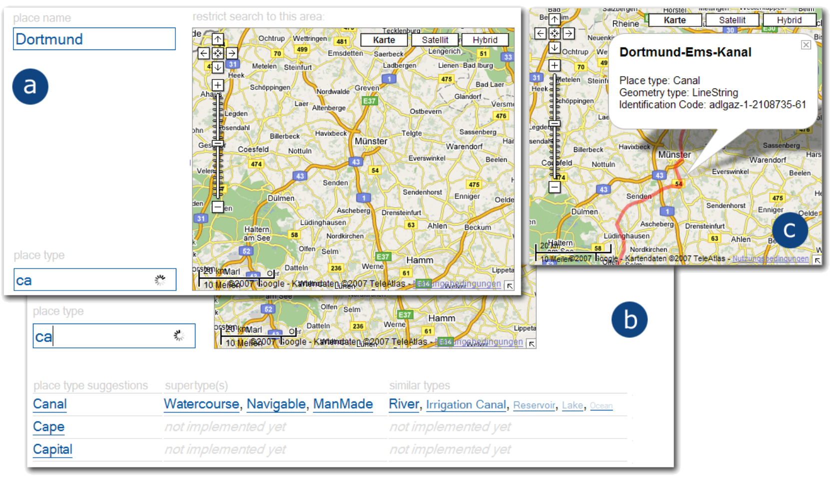

Figure 10: Conceptual design for the Web interface: search interface with input fields for place name and type, and a map for spatial restriction (a); automatic suggestion of place types during user input (b); display of results as map overlay (c) (based on Janowicz and Keßler [83]).

The proposed interface makes use of the AJAX[7] technology to create a *search-while-you-type* input field: as the user enters the place type, results are automatically loaded in the background. The suggested types are based on a syntactic match of the letters already typed in by the user. Next, to every suggestion, its supertypes and the most similar types from the ontology are presented [87]. The font size of the displayed types indicate their similarity to the suggested type in the leftmost column. This technique is named font-size scaling or tag cloud

7 Asynchronous JavaScript and XML

(algorithm). This way, there is no need for the user to know about the underlying feature type hierarchy, as similar types are automatically suggested by the interface. All suggestions are hyperlinked and can be moved to the input field with a single click. The interface also allows for spatial restriction by zooming the map to the desired extent. The proposed interface thus allows for an intuitive workflow that supports also novice users in the selection of the appropriate feature types for a query. Apart from up-to-date Web technology implemented as so called *mashup* (including the map view provided by Google Maps™), this functionality is made possible by the feature type ontology [83] in the background, and by the SIM-DL similarity server accessing it [85].

The semantics-enabled gazetteer Web interface is under development as part of the SimCat project, focusing on the mapping between the feature type ontology and the feature categorization provided by ADL (using the feature type thesaurus).

5 SIM-DL SIMILARITY THEORY

While section 2.2 describes the process of measuring similarity on an abstract level, this chapter presents the SIM-DL theory based on the introduced framework. Besides abstract specifications using generic symbols such as C and D, several examples from hydrology (and Baader's classical family ontology [6]) are given to illustrate how SIM-DL works. In contrast to other DL or fuzzy set based theories (e.g., [26, 32, 33, 177]), SIM-DL works exclusively on the concept (i.e., terminological) level. No knowledge about individuals (via assertions) is taken into account. To make this point clear, section 5.7 compares SIM-DL to related approaches.

5.1 A FORMAL ACCOUNT OF SIMILARITY

The following sections specify how SIM-DL compares concepts for similarity. Besides plain text descriptions, the processing steps and similarity functions are motivated based on the interpretation of the compared DL concepts. As set theory is usually presented as an axiomatized theory expressed using predicate logic; similarity functions are characterized in terms of first order logic. For reasons of readability, an abbreviated syntax is used which allows to embed symbols for exclusive OR ($\oplus$) and element-of ($\in$) within definitions.

The logical foundation underlying SIM-DL needs to be clearly separated from both the implementation and the later application. In the following, similarity is modeled as class level relation between compared concepts. Three relationships are distinguished: $\simeq$, $\gtrless$ and $\nsim$. The relation $\simeq$ holds between concepts that are either equal or not distinguishable (from a similarity-based point of view). $\gtrless$ holds between concepts that share at least some superconcepts while others are different. Finally, $\nsim$ stands for not similar – indicating that compared concepts have no superconcept in common. These relationships can be mapped to similarity values, $\simeq$ corresponds to a value of 1, $\gtrless$ to a value between 0 and 1 (]0,1[), while $\nsim$ maps to a similarity value of 0.

For all language constructors, the conditions under which concept comparison yields a value of 1,]0,1[, and 0 are defined. Based on these definitions, functions are introduced which specify how similarity is computed. While the logical framework only distinguishes between 0,]0,1[, and 1, the functions map compared concept descriptions to real numbers $\mathbb{R}_{[0,1]}$. A similarity function that satisfies the logical foundations is called a valid realization. In dependence of the later application area, there may be more than one appropriate function (which returns slightly different similarity values). The implementation of SIM-DL within the similarity server is discussed in chapter 6.

To illustrate how similarity functions are defined, consider the following example. The expression $sim(C, D) = 1$ is a statement in the domain of similarity values ($\mathbb{R}_{[0,1]}$). It maps from two concepts to a similarity value. We assume that the interpretation of these values is consistent with the interpretation of real numbers[1], i.e., if $sim(C, D) = 1$ and $sim(C, E) = 0.5$ then D is more similar to C than E (to C). Next, it has to be specified under which conditions the similarity function maps C and D to 1. This is achieved by definitions which map to *true* or *false* (e.g., $\simeq (C^{\mathcal{I}}, D^{\mathcal{I}})\ iff ...$). If $\simeq (C^{\mathcal{I}}, D^{\mathcal{I}})$ is *true*, $sim(C, D) = 1$, else $sim(C, D) \neq 1$. As one can see, the FOL definitions are specified on sets (i.e., the interpretation of DL concepts). It is also possible to interpret these sets as individuals in FOL, which also allows to model DL roles. The formal semantics of description logics maps $(.^{\mathcal{I}})$ to set theory which (as argued above) can be expressed in terms of FOL. Keeping this relationship between first order axiomatization and similarity functions in mind, one can state that $sim(C, D) = 1$ iff $\simeq (C^{\mathcal{I}}, D^{\mathcal{I}})$.

5.2 CONTEXT, SEARCH AND TARGET CONCEPTS

In SIM-DL the search concept (C_s) can be either selected from the used ontology or phrased using primitives, roles, and concepts defined within the ontology (see figure 8). While it is possible to select a single target concept (C_t) by hand, the context-based (C_c) target selection is used by default in SIM-DL. This corresponds to the third case specified in the framework (see section 2.2.1). This is for two reasons. First, within SIM-DL context does not only influence which concepts are compared but also how similar they are (see section 5.5 and especially 5.5.10 for more details on context sensitivity). Second, a single similarity value is difficult to interpret.

A similarity value (e.g., 0.67) computed between two concepts hides most of the important information. It does neither 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. Imagine an ontology where the least similar target concept has a value of 0.6 (compared to the search concept), while the comparison to the most similar concept yields 0.9. In such case, a similarity value of 0.67 is not high at all. Besides these interpretation problems, isolated comparison puts to much stress on the concrete similarity value. It is hard to argue that and why the result is (cognitively) plausible without other reference values [89].

Instead, SIM-DL focuses on similarity rankings. The search concept is compared to all target concepts derived from the context. The result is an ordered list with descending similarity values. In this work, we do not argue that similarity values are cognitively plausible, but that the computed order correlates with human ranking judgments (see chapter 7). One can argue that such rating puts a single similarity value in context (namely a result context), however this term is not used here to avoid confusion.

SIM-DL distinguishes between two kinds of context, the internal, and the external context. The internal context[2] corresponds to the con-

1 This assumption is discussed in chapter 8 in more detail. One could also argue that "1" is a symbol ($\bar{1}$) from the domain of similarity which is interpreted as number 1 from the domain of real numbers.

2 Which we will refer to by *context*.

text definition in the similarity framework. Along with MDSM [143], context is defined as the sets of those (target) concepts which are subconcepts of the context concept $\{C_t | C_t \sqsubseteq C_c\}$. Again, the context concept can be either part of the examined ontology or phrased by the user. In contrast to the internal context, the external context is neither part of the examined ontology nor the similarity theory. The external context can be thought of as a set of rules that specify which descriptors of the selected target concepts are relevant and which should be left aside. In terms of geographic feature types, whether *Forest* and *Park* are similar also depends on spatial and temporal aspects. In summer, forests are threatened by fire which should be reflected in the conceptualization of forests and therefore influence similarity [92]. As the external context has only indirect impact on similarity (by modifying the compared concepts), it is not discussed here. Further details about kinds of contexts are discussed in chapter 8.

5.3 CANONICAL NORMAL FORM

Before similarity can be computed, the compared (complex) concepts have to be rephrased to the following $\mathcal{ALCHQ}$ disjunctive normal form (DNF): A concept description C is in normal form, iff $C = \top$, $C = \bot$ or $C = C_1 \sqcup ... \sqcup C_n$ and each $C_i (i = 1, ...n)$ is of the form:

$$C := \bigsqcap_{A \in primitive(C_i)} A \sqcap \bigsqcap_{R \in N_R} \left(\bigsqcap_{C' \in exists_R(C_i)} (\exists R.C') \sqcap \forall R.forall_R(C_i) \right. \\ \left. \sqcap \bigsqcap_{C' \in min_R(C_i)} (\geq |min_R(C_i)| R.C') \sqcap \bigsqcap_{C' \in max_R(C_i)} (\leq |max_R(C_i)| R.C') \right) \tag{5.1}$$

The set $primitive(C)$ represents all (negated) primitives (and $\bot$) at the top-level of C. N_R is the set of available roles, and $exists_R(C)$, $min_R(C_i)$, and $max_R(C_i)$ denote the sets of all C' for which there exists $\exists R.C'$ (min/max restrictions, respectively) on the top-level of C. $forall_R(C_i)$ denotes the intersection of concepts $(C1 \sqcap ... \sqcap C_n)$ derived by merging all value restrictions for the role R ($\forall R.C_i$) on the top level of C. $|min_R(C_i)|$ and $|max_R(C_i)|$ represent the minimum and maximum cardinalities for the role R on the top-level of C. Note that the concepts $forall_R(C_i)$ and C' are again in $\mathcal{ALCHQ}$ normal form.

Basically, the normal form is derived by unfolding (also called expanding) concepts [72, 124], sorting their descriptors, and applying De Morgan's laws. In case of equality ($\equiv$), unfolding is trivial because all non-primitive concept names are replaced by their definitions which are then recursively unfolded. If a concept C is described using a concept D ($D \equiv A$), each occurrence of D in C is replaced by its definition, i.e., A. If A is non-primitive, the process is recursively applied to A (its superconcepts). For inclusion axioms such as $D \sqsubseteq A$, a placeholder concept D' needs to be introduced. This placeholder represents the *primitiveness* [72] of D that distinguishes it from A. One has to keep in mind that in case of inclusion it is not specified how D differs from A. In terms of the example used above, each occurrence of D in C is replaced by $D' \sqcap A$ while A is recursively unfolded.

To ensure that the SIM-DL measure is not influenced by its syntactic form, rewriting rules (see also [20, 72, 124]) have to be applied to get a canonical representation of the compared concepts. This step is also sometimes referred to as simplification. On the one hand these rewriting rules map between equivalent expressions such as $(\forall R.\bot)$ and $(\leq 0R.\top)$. On the other hand they ensure that only such descriptions are used within concept specifications which (by definition) have an impact on the cardinality of the regarded sets (i.e., the interpretation). For instance, $(\geq 1R.C) \sqcap (\geq 2R.C)$ is mapped to $(\geq 2R.C)$, while $(... \sqcap \top)$ can be skipped without changing the extent of the specified concept.

Three groups of rules can be distinguished, satisfiability rules, merging rules, and uniformity rules. The first group plays a major role in subsumption reasoning as these rules reduce the effort of satisfiability checking by filtering obvious inconsistencies. For the sake of completeness, and because SIM-DL requires subsumption reasoning (see section 5.4), some of these rules are listed here. For the rest of this chapter we assume the TBox $\mathcal{T}$ to be consistent, i.e., all concepts and roles are satisfiable. In other words, similarity is only measured between contradiction-free concepts. The second group consists of rules that filter and merge redundant descriptors (as described above for number restrictions). While the first groups can be applied to all concepts separately, the last group of rules depends on the search and target concept. This group consists of normalization rules intended to make two concepts comparable, i.e., base them on the same constructor. Besides classical deduction (inference) rules such De Morgan's laws or Disjunctive Syllogism are not listed here, the following rewriting rules are applied to concepts in $\mathcal{ALCHQ}$ normal form.

Rewriting Rule 5.1 (complex negation)
a) *Condition:* $\neg C$ *where* $C = \top$
Action: Rewrite $\neg C$ *to* $\bot$
b) *Condition:* $\neg C$ *where* $C = \bot$
Action: Rewrite $\neg C$ *to* $\top$
c) *Condition:* $\neg C$ *where* $C = \neg D$
Action: Rewrite $\neg C$ *to* D

Rewriting Rule 5.2 (intersection)
a) *Condition:* $... \sqcap C \sqcap C'$ *where* $C' = \top$
Action: Rewrite $... \sqcap C \sqcap C'$ *to* $... \sqcap C$
b) *Condition:* $... \sqcap C \sqcap C'$ *where* $C' = \bot$
Action: Rewrite $... \sqcap C \sqcap C'$ *to* $\bot$
c) *Condition:* $... \sqcap C \sqcap C'$ *where* $C' = \neg C$
Action: Rewrite $... \sqcap C \sqcap C'$ *to* $\bot$
d) *Condition:* $... \sqcap C \sqcap C'$ *where* $C' \sqsubseteq C$
Action: Rewrite $... \sqcap C \sqcap C'$ *to* $... \sqcap C'$

Rewriting Rule 5.3 (union)
a) *Condition:* $... \sqcup C \sqcup C'$ *where* $C' = \top$
Action: Rewrite $... \sqcup C \sqcup C'$ *to* $\top$
b) *Condition:* $... \sqcup C \sqcup C'$ *where* $C' = \bot$
Action: Rewrite $... \sqcup C \sqcup C'$ *to* $... \sqcup C$

Rewriting Rule 5.4 (value restriction)
a) *Condition:* $\forall R.C \sqcap \forall R.C'$
Action: Rewrite $\forall R.C \sqcap \forall R.C'$ *to* $\forall R.(C \sqcap C')$
b) *Condition:* $\forall R.C \sqcup \forall R.C'$
Action: Rewrite $\forall R.C \sqcup \forall R.C'$ *to* $\forall R.(C \sqcup C')$
c) *Condition:* $\forall R.C$ *where* $C = \top$
Action: Rewrite $\forall R.C$ *to* $\top$

Rewriting Rule 5.5 (existential quantification)
a) *Condition:* C_s *contains* $\exists R.C$ *while* C_t *contains* $(\geq nR.C)$ *and* $n \geq 0$
Action: Rewrite $(\geq nR.C)$ *to* $\exists R.C$

Rewriting Rule 5.6 (qualified number restrictions)
a) *Condition:* C_s *contains* $(\geq nR.C)$ *while* C_t *contains* $\exists R.C$ $n \geq 0$
Action: Rewrite $\exists R.C$ *to* $(\geq nR.C)$
b) *Condition:* $(\geq nR.C)$ *and* $n = 0$.
Action: Rewrite $(\geq nR.C))$ *to* $\top$
c) *Condition:* $(\leq nR.C)$ *and* $n \leq 0$.
Action: Rewrite $(\leq nR.C))$ *to* $\perp$

Note that the described rules are recursively applied and combined. For instance, $\forall R.(C \sqcap \neg C)$ is simplified to $\forall R.\perp$ and finally (if necessary) to $(\leq 0R.\top)$. Similarly, the complex concept $\forall R.C \sqcap \forall R.\neg C$ is simplified to $\perp$ using the rewriting rule 5.2c. The rule 5.2d is of special importance, as it allows for extended rules such as:

Rewriting Rule 5.7 (exemplarily derived rules)
a) *Condition:* $(\geq nR.C) \sqcap (\geq mR.C)$ *and* $n \geq m$.
Action: Rewrite $(\geq nR.C) \sqcap (\geq mR.C)$ *to* $(\geq nR.C)$
b) *Condition:* $(\leq nR.C) \sqcap (\leq mR.C)$ *and* $n \leq m$.
Action: Rewrite $(\leq nR.C) \sqcap (\leq mR.C)$ *to* $(\leq nR.C)$

Besides rewriting rules used for simplification, one may also think of rules which add concept descriptions. This is important in case of negation. A simplified example can be constructed as follows. The search concept is defined as $C_s \equiv A \sqcap D \sqcap \neg E$ and the target concept is specified as $C_t \equiv A \sqcap \neg D$. If we assume that $E \sqsubseteq D$, it follows that C_t is also $\neg E$. Instead of one common and two distinctive superclasses, C_s and C_t share two common superclasses while one is different.

The rewriting rules are language specific and have to be modified if other description logics than $\mathcal{ALCHQ}$ are used.

5.4 ALIGNMENT MATRIX

To compare concepts using SIM-DL, an alignment matrix M_1 with all possible combinations of their descriptors is created. Out of all resulting tuples, those with the highest similarity values (see section 5.5) are chosen for further computation. Each descriptor from C_s and C_t, respectively, is only selected once. Similarity cannot be measured between descriptors that are based on different constructors, i.e., the resulting similarity is always 0. In case of quantifications and restrictions similarity is derived from comparing roles and fillers. Therefore additional alignment matrices M_n have to be created if the fillers are complex concepts again (see figure 11).

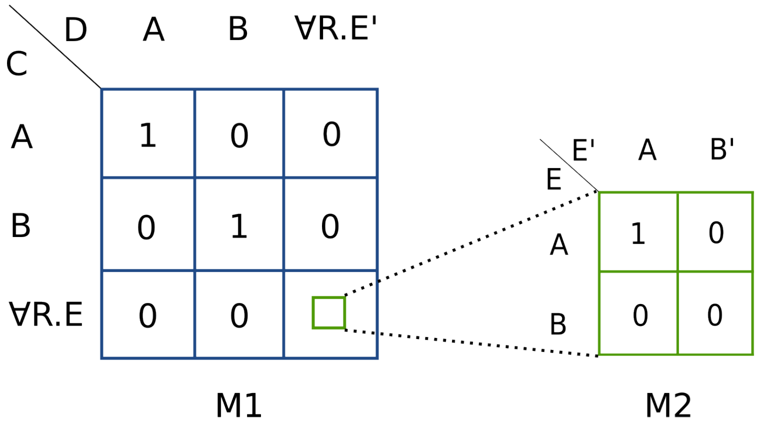

Figure 11: Alignment Matrices M1 and M2 for sim(C,D).

Each matrix takes care of its own normalization, i.e., the similarity value for a matrix is always between 0 and 1. To achieve this, the normalization factor is increased by 1 for each selected tuple of concept descriptors from C_s and C_t. In terms of the simplified example depicted in figure 11, the resulting similarity is $\frac{2.5}{3}$ (0.83)[3].

In contrast to some findings from psychology (see [56, p.18] for an overview), in SIM-DL we assume self similarity for all concepts and roles, i.e., *sim(C,C)* (respectively *sim(R,R)*) is always 1. All concepts are equally similar to themselves. The other way around, one cannot infer from a similarity value of 1 that compared concepts (*C* and *D*) are equal (in terms of identity), but that there is no meaningful distinction between them in the particular application area.

As in most feature and geometric approaches, SIM-DL is an asymmetric measure, i.e., $sim(C_s, C_t)$ is not necessarily equal to $sim(C_t, C_s)$. The comparison of two concepts depends therefore not only on their descriptors but also on the direction in which both are compared. If C_s is defined by more descriptors than C_t, the similarity value for each of these descriptors is set to 0, and the normalization factor is increased by 1. In the opposite case, the normalization factor is not increased. In addition, the similarity betwen C_s and C_t (or their descriptors) is always 1 if C_t is a subconcept of C_s. The other way around, the similarity $sim(C_s, C_t)$ is less than 1 (in most cases[4]) if C_t is a superconcept of C_s.

SIM-DL serves information retrieval tasks with a clear separation of search and target concept. As described in the use case chapter 4, the user selects a concept using a search-as-you-type interface instead of creating a query concept. The gazetteer Web interface returns all direct superconcepts and a list of similar concepts. In such setting, the search concept is the referent of the query while the compared-to concepts are the variants [168]. This corresponds to a combination of the first and second scenario discussed for alignment in the framework

3 $sim(C, D) = \frac{sim(A,A)+sim(B,B)+\left(sim(R,R)*\frac{sim(A,A)+sim(B,B')}{2}\right)}{3} = 0.83$; see section 5.5.

4 See section 5.5.4 for details about how normalization is handled in case of the union constructor.

(see section 2.2.3). All subconcepts of the search concept satisfy the user's requirements [111]. In contrast, the similarity section lists those concepts for which some requested descriptors are missing or differ from those of the search concept; see figure 10.

At first sight, this approach contradicts the notion of asymmetry introduced in Tversky's classical paper about features of similarity [168]:

> Similarity judgments can be regarded as extensions of similarity statements, that is, statements of the form "a is like b." Such a statement is directional; it has a subject, a, and a referent, b, and it is not equivalent in general to the converse similarity statement "b is like a." In fact, the choice of subject and referent depends, at least in part, on the relative salience of the objects. We tend to select the more salient stimulus, or the prototype, as a referent, and the less salient stimulus, or the variant, as a subject. [...] We say "an ellipse is like a circle," not "a circle is like an ellipse," and we say "North Korea is like Red China" rather than "Red China is like North Korea."
>
> As will be demonstrated later, this asymmetry in the choice of similarity statements is associated with asymmetry in judgments of similarity. Thus, the judged similarity of North Korea to Red China exceeds the judged similarity of Red China to North Korea. Likewise, an ellipse is more similar to a circle than a circle is to an ellipse. Apparently, the direction of asymmetry is determined by the relative salience of the stimuli; the variant is more similar to the prototype than vice versa. [168, p. 335]

This apparent contradiction has the following reasons. First of all, Tversky's examples focus on individuals rather than concepts (though his theory was extended to concepts as well). Statements such as "a is like b", even if applied to concepts, do not necessarily imply a search direction. North Korea is more similar to Red China because the salient aspect chosen for comparison is the political system. In this case Red China is the prominent stimulus (i.e., prototype) [57, 115, 123, 158]. In case of SIM-DL, the compared conceptualizations are models rather than prototypes and the compared descriptors are determined by the search concept. If a user is searching for navigable waterways, then navigable and man-made waterways satisfy these needs. This is not the case, if the user explicitly searches for entities which are also man-made, while the retrieved entities are not.

While this setting meets the requirements of the gazetteer use case, other information retrieval tasks may require a symmetric similarity measure where every difference between C_s and C_t reduces similarity. Such a measure can act as an extension to the subsumption-based retrieval methodology presented by Lutz and Klien [111]. Table 2 shows the impact of asymmetry on similarity using the example described above; where:

- $C \equiv Waterway \sqcap Navigable$
- $D \equiv Waterway \sqcap Navigable \sqcap ManMade$

- $E \equiv Waterway \sqcap NavigableSmallVessels \sqcap ManMade$
- $NavigableSmallVessels \sqsubseteq Navigable$

Table 2: The impact of asymmetry on similarity.

Symmetry	Search Direction	Similarity	Subsumption
Asymmetric	sim(C,D)	1	$\surd$
Asymmetric	sim(D,C)	0.67	$\times$
Asymmetric	sim(C,E)	1	$\surd$
Asymmetric	sim(E,C)	0.5	$\times$
Symmetric	sim(C,D)	0.67	$\surd$
Symmetric	sim(D,C)	0.67	$\times$
Symmetric	sim(C,E)	0.5	$\surd$
Symmetric	sim(E,C)	0.5	$\times$

If C is the query concept, the subsumption-based retrieval approach would list (as opposed to the symmetric similarity approach) D and E as tantamount results, without pointing out that E is not navigable for same kinds of vessels.

5.5 SIMILARITY FUNCTIONS

This section describes the similarity functions necessary to compare concepts specified using the $\mathcal{ALCHQ}$ language. SIM-DL defines similarity functions for each available constructor. The measurement process always starts at the union level (see $\mathcal{ALCHQ}$ canonical normal form; section 5.3) with the sim_u function. All concepts at this level are formed by intersection and their similarity is determined by sim_i. The intersected concepts are either primitives (sim_p), existential quantifications (sim_e), value restrictions (sim_f), or qualified number restrictions (sim_{min} or sim_{max}, respectively). In addition to role hierarchies (sim_r), SIM-DL supports temporal and topological neighborhood models (sim_n) to compute similarity between roles. This allows to determine similarity between tuples such as $(\exists inside.Lake, \exists overlap.Lake)$.

The similarity functions are specified in such a way that they can be used for both symmetric and asymmetric measures (see section 5.4) without major modifications. To improve readability, the following equations are abstracted from their concrete implementation and sequence within the similarity server. Their integration into the SIM-DL server is described in chapter 6 and appendix A. Additional figures are used to illustrate how particular similarity functions work. Measures between complex concepts (involving roles) are difficult to depict, therefore the figures are examples rather than replacements of textual definitions.

5.5.1 *Primitive Concepts*

Each named concept from the set $\mathcal{N}_\mathcal{T}$ is expressed in terms of primitives forming the base symbols $\mathcal{B}_\mathcal{T}$ of the considered terminology $\mathcal{T}$.

Such primitives occur only on the right-hand side of axioms, i.e., they have no definitions to be compared.

To measure similarity between primitives (sim_p), an adapted version of the Jaccard Similarity Coefficient is used. The (binary) Jaccard coefficient measures the degree of overlap between two sets as the ratio of the cardinality of shared attributes (e.g., features) from $S_1 \wedge S_2$ to the cardinality retrieved from $S_1 \vee S_2$. Rodríguez and Egenhofer [143] use an asymmetric version of Jaccard's coefficient within their Matching Distance Similarity Measure (see also [168]). In SIM-DL, the Jaccard Similarity Coefficient is adopted to compute the normalized and context-aware co-occurrence of primitives within the definitions of other (non-primitive) concepts. Two primitives are the more similar, the more complex concepts are defined using both (and not only one of them). If $sim_p(A, B) = 1$, both primitives always co-occur in complex concepts and therefore cannot be distinguished.

Definition 5.1 *The comparison of two primitives A and B yields 1, if every concept C (out of the context set) specified using the primitive A is also defined by using B and vice versa. In other words, if there is no such concept C that is either defined by A or B but not by both (see figure 12a).*

Given $A, B \in \mathcal{B}_\mathcal{T}$:
$\simeq_p (A^\mathcal{I}, B^\mathcal{I})$ *iff* $\forall C^\mathcal{I}[(C^\mathcal{I} \subseteq (A^\mathcal{I} \cap {C_c}^\mathcal{I})) \leftrightarrow (C^\mathcal{I} \subseteq (B^\mathcal{I} \cap {C_c}^\mathcal{I}))]$

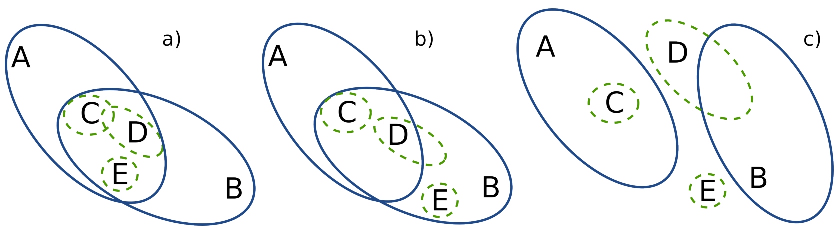

Figure 12: Set-based visualization of similarity between primitive concepts: a) $sim_p(A, B) = 1$, b) $sim_p(A, B) =]0, 1[$, and c) $sim_p(A, B) = 0$.

Definition 5.2 *The comparison of two primitives A and B yields a value between 0 and 1, if at least one concept C (out of the context set) specified using the primitive A is also defined by using B (and vice versa), and there is at least one C′ that is either defined using A or B but not by both (see figure 12b).*

Given $A, B \in \mathcal{B}_\mathcal{T}$:
$\lesssim_p (A^\mathcal{I}, B^\mathcal{I})$ *iff* $\exists C^\mathcal{I} \exists C'^\mathcal{I}[(C^\mathcal{I} \subseteq (A^\mathcal{I} \cap B^\mathcal{I} \cap {C_c}^\mathcal{I}))$
$\wedge((C'^\mathcal{I} \subseteq (A^\mathcal{I} \cap {C_c}^\mathcal{I})) \oplus (C'^\mathcal{I} \subseteq (B^\mathcal{I} \cap {C_c}^\mathcal{I})))]$

Definition 5.3 *The comparison of two primitives A and B yields 0, if no concepts C (out of the context set) defined using the primitive A is also defined by B (and vice versa). In other words there is no such concept C that is at the same time defined using A and B (see figure 12c).*

Given $A, B \in \mathcal{B}_\mathcal{T}$:
$\nsim_p (A^\mathcal{I}, B^\mathcal{I})$ *iff* $\neg\exists C^\mathcal{I}[C^\mathcal{I} \subseteq (A^\mathcal{I} \cap B^\mathcal{I} \cap {C_c}^\mathcal{I})]$

REALIZATION Based on the definitions 5.1, 5.2, and 5.3, SIM-DL implements similarity between primitives as the ratio of the cardinality of common concepts specified using A and B to the cardinality of distinguishing concepts (see equation 5.2).

$$sim_p(A,B) = \frac{|\{C \mid (C \sqsubset A) \wedge (C \sqsubset B)\}|}{|\{C \mid (C \sqsubset A) \vee (C \sqsubset B)\}|} \tag{5.2}$$

5.5.2 *Role Hierarchy*

The language $\mathcal{ALCHQ}$ (see section 3.2.1) supports role hierarchies, i.e., role inclusion, but does not allow for role constructors such as intersection (see [79, 80] for a possible extension of SIM-DL), union, complement, or composition. Same as argued for primitives, there are no role definitions which can be compared for similarity. Because of the missing intersection constructor we do not apply Jaccard's Coefficient here. Instead, a network-based approach [137] is taken to compute the similarity of roles (R and S) within a role hierarchy. Similarity (sim_r) is expressed as the ratio between the shortest path from R to S and the maximum path within the graph representation of the role hierarchy.

To explain how sim_r is specified in first order logic, one needs to introduce the (DL) universal role U and its interpretation $U \equiv \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. As described in section 3.2.1, each role R is interpreted as $R \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ and hence is a subrole of U.

Note that, compared to sim_p, similarity between roles is defined without reference to the context. This would require to take only such roles T (see below) into account which are used within quantifications or restrictions of concepts within the context.

Definition 5.4 *The comparison of two roles R and S yields 1, if R is a superrole of S. (see figure 13a).*

Given $R, S \in \mathcal{N}_{\mathcal{R}}$:
$\simeq_r (R^{\mathcal{I}}, S^{\mathcal{I}})$ *iff* $(R^{\mathcal{I}} \supseteq S^{\mathcal{I}})$

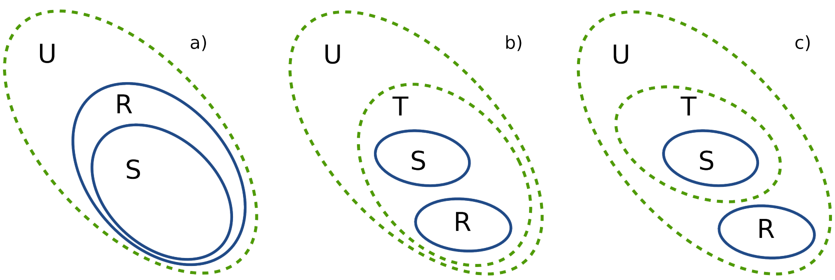

Figure 13: Set-based visualization of similarity between roles:
a) $sim_r(R,S) = 1$, b) $sim_r(R,S) =]0,1[$, and c) $sim_r(R,S) = 0$.

Definition 5.5 *The comparison of two roles R and S yields a value between 0 and 1, if there exists at least one subrole T of U which at the same time is a superrole of R and S (but R is not a subrole of S; see figure 13b).*

Given R, S and $T \in \mathcal{N}_{\mathcal{R}}$:
$\lesssim_r (R^{\mathcal{I}}, S^{\mathcal{I}})$ *iff* $\exists T^{\mathcal{I}}[(T^{\mathcal{I}} \subset U^{\mathcal{I}}) \wedge (R^{\mathcal{I}} \subseteq T^{\mathcal{I}}) \wedge (S^{\mathcal{I}} \subseteq T^{\mathcal{I}})] \wedge (R^{\mathcal{I}} \not\supseteq S^{\mathcal{I}})$

Definition 5.6 *The comparison of two roles R and S yields 0, if they do not share a common superrole T ($T \sqsubset U$) and R is not a subrole of S (see figure 13c).*

Given $R, S \in \mathcal{N}_{\mathcal{R}}$:
$\nsim_r (R^{\mathcal{I}}, S^{\mathcal{I}})$ *iff* $\neg\exists T^{\mathcal{I}}[(T^{\mathcal{I}} \subset U^{\mathcal{I}}) \wedge (R^{\mathcal{I}} \subseteq T^{\mathcal{I}}) \wedge (S^{\mathcal{I}} \subseteq T^{\mathcal{I}})]$
$\wedge (R^{\mathcal{I}} \nsubseteq S^{\mathcal{I}}) \wedge (S^{\mathcal{I}} \nsubseteq R^{\mathcal{I}})$

REALIZATION Based on the definitions 5.4, 5.5, and 5.6, similarity between roles (sim_r) is their normalized distance within the graph representation of the role hierarchy. The normalization is depth-dependent to indicate that the distance from node to node decreases with increasing depth level of R and S within the hierarchy. In other words, the weights of the edges used to determine the path between R and S decrease with increasing depth of the graph. If a path between two roles crosses U, similarity is 0.

$$sim_r(R, S) = \frac{depth(lub(R, S))}{depth(lub(R, S)) + edge\_distance(R, S)} \tag{5.3}$$

5.5.3 *Intersection*

A concept C can be described by the intersection of two other concepts A and B, i.e., $C^{\mathcal{I}} \subseteq (A^{\mathcal{I}} \cap B^{\mathcal{I}})$. These superconcepts are either composed concepts themselves or primitives. In the first case, the concepts can be further decomposed up to a level where only primitive concepts (as well as those formed by quantifications and restrictions) are left.

In terms of similarity, the intersection constructor can be interpreted as a normalized summation function. If the intersected superconcepts forming C and D are primitives, similarity between C and D is derived from the similarity values of $sim_p(A, A')$ and $sim_p(B, B')$. Else, if the compared concepts are formed by complex concepts again, similarity is derived by unfolding these concepts. The same procedure is applied if the compared concepts are intersections of more than two concepts.

Definition 5.7 *The comparison of two (complex) concepts C and D formed by intersection yields 1, if the similarity of all their superconcepts (forming the intersection) yields 1. If the superconcepts are primitives, $sim_i(C, D) = 1$ if $sim_p(A, A')$ and $sim_p(B, B')$ are 1 (see figure 14a).*

Given $A, A', B, B' \in \mathcal{B}_{\mathcal{T}}$ and $C \equiv A \sqcap B$ and $D \equiv A' \sqcap B'$:
$\simeq_i (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $\simeq_p (A^{\mathcal{I}}, A'^{\mathcal{I}}) \wedge \simeq_p (B^{\mathcal{I}}, B'^{\mathcal{I}})$

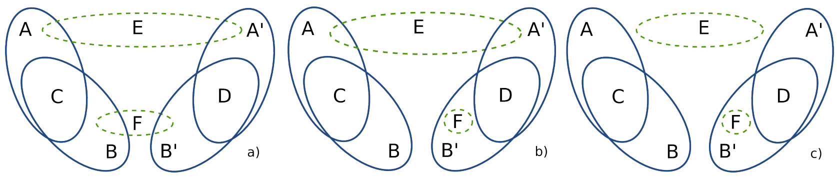

Figure 14: Set-based visualization of similarity between concepts formed by intersection: a) $sim_i(C, D) = 1$, b) $sim_i(C, D) =]0, 1[$, and c) $sim_i(C, D) = 0$.

Definition 5.8 *The comparison of two (complex) concepts C and D formed by intersection yields a value between 0 and 1, if at least for one pair of compared superconcepts the resulting similarity is not 1 (while it is not 0 for all)(see figure 14b).*

Given $A, A', B, B' \in \mathcal{B}_\mathcal{T}$ *and* $C \equiv A \sqcap B$ *and* $D \equiv A' \sqcap B'$:
$\lessgtr_i (C^\mathcal{I}, D^\mathcal{I})$ *iff* $(\lessgtr_p (A^\mathcal{I}, A'^\mathcal{I}) \vee \lessgtr_p (B^\mathcal{I}, B'^\mathcal{I}))$
$\vee(\simeq_p (A^\mathcal{I}, A'^\mathcal{I}) \oplus \simeq_p (B^\mathcal{I}, B'^\mathcal{I}))$

Definition 5.9 *The comparison of two (complex) concepts C and D formed by intersection yields 0, if the similarity of all their superconcepts (forming the intersection) yields 0. In case where these superconcepts are primitives,* $sim_i(C, C') = 0$ *if* $sim_p(A, A')$ *and* $sim_p(B, B')$ *are 0 (see figure 14c).*

Given $A, A', B, B' \in \mathcal{B}_\mathcal{T}$ *and* $C \equiv A \sqcap B$ *and* $D \equiv A' \sqcap B'$:
$\nsim_i (C^\mathcal{I}, D^\mathcal{I})$ *iff* $\nsim_p (A^\mathcal{I}, A'^\mathcal{I}) \wedge \nsim_p (B^\mathcal{I}, B'^\mathcal{I})$

REALIZATION Based on the definitions 5.7, 5.8, and 5.9, and following the $\mathcal{ALCHQ}$ canonical normal form (see section 5.3), each C_i (and D_j, respectively) is an intersection of primitives or concepts formed by restrictions or quantifications. sim_i is specified as the normalized sum of the similarity values computed for the involved tuples *(X,Y)* (see equation 5.4). The normalization factor σ corresponds to the number of these tuples. Consequently the possible results of sim_i range between 0 and 1. S_i is the set of selected tuples on the intersection level.

$$sim_i(C, D) = \frac{1}{\sigma} \sum_{(X,Y) \in S_i} sim(X, Y) \qquad (5.4)$$

5.5.4 *Union*

A concept C can be derived by the union of two concepts E and F, i.e., $C^\mathcal{I} \subseteq (E^\mathcal{I} \cup F^\mathcal{I})$. These concepts are either composed concepts or primitives. In the first case, the concepts can be further decomposed up to a level where only primitive concepts (as well as those formed by quantifications and restrictions) are left. The same procedure is applied if the compared concepts are unions of more than two concepts.

Following the $\mathcal{ALCHQ}$ normal form (see section 5.3), the comparison of two concepts always starts with sim_u. In terms of similarity, the union constructor can be either interpreted as the weighted sum or as the maximum (respectively minimum) similarity of the involved tuples.

Definition 5.10 *The comparison of two (complex) concepts C and D formed by union yields 1, if the similarity of at least one of their subconcepts (forming the union) yields 1. In case where these concepts are primitives,* $sim_u(C, D) = 1$ *if* $sim_p(A, A')$ *(inclusive) or* $sim_p(B, B')$ *is 1 (see figure 15a).*

Given $A, A', B, B' \in \mathcal{B_T}$ *and* $C \equiv A \sqcup B$ *and* $D \equiv A' \sqcup B'$:
$\simeq_u (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $\simeq_p (A^{\mathcal{I}}, A'^{\mathcal{I}}) \vee \simeq_p (B^{\mathcal{I}}, B'^{\mathcal{I}})$

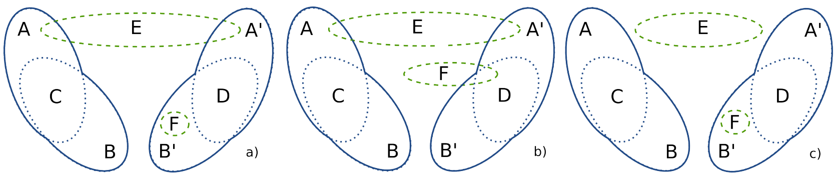

Figure 15: Set-based visualization of similarity between concepts formed by union: a) $sim_u(C, D) = 1$, b) $sim_u(C, D) =]0, 1[$, and c) $sim_u(C, D) = 0$.

Definition 5.11 *The comparison of two (complex) concepts C and D formed by union yields a value between 0 and 1, if at least for one pair of compared subconcepts the resulting similarity is not 0, and there is no such pair for which similarity is 1 (see figure 15b).*

Given $A, A', B, B' \in \mathcal{B_T}$ *and* $C \equiv A \sqcup B$ *and* $D \equiv A' \sqcup B'$:

$$\begin{aligned} \lessgtr_u (C^{\mathcal{I}}, D^{\mathcal{I}}) \text{ iff } & (\lessgtr_p (A^{\mathcal{I}}, A'^{\mathcal{I}}) \wedge \lessgtr_p (B^{\mathcal{I}}, B'^{\mathcal{I}})) \\ & \vee (\nsim_p (A^{\mathcal{I}}, A'^{\mathcal{I}}) \wedge \lessgtr_p (B^{\mathcal{I}}, B'^{\mathcal{I}})) \\ & \vee (\nsim_p (B^{\mathcal{I}}, B'^{\mathcal{I}}) \wedge \lessgtr_p (A^{\mathcal{I}}, A'^{\mathcal{I}})) \end{aligned}$$

Definition 5.12 *The comparison of two (complex) concepts C and D formed by union yields 0, if the similarity of all their subconcepts (forming the union) yields 0. In case where the superconcepts are primitives,* $sim_i(C, C') = 0$ *if* $sim_p(A, A')$ *and* $sim_p(B, B')$ *are 0 (see figure 15c).*

Given $A, A', B, B' \in \mathcal{B_T}$ *and* $C \equiv A \sqcup B$ *and* $D \equiv A' \sqcup B'$:
$\nsim_u (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $\nsim_p (A^{\mathcal{I}}, A'^{\mathcal{I}}) \wedge \nsim_p (B^{\mathcal{I}}, B'^{\mathcal{I}})$

REALIZATION Based on the definitions 5.10, 5.11, and 5.12, and following the $\mathcal{ALCHQ}$ canonical normal form, sim_u can be implemented in two ways. One can either decide to compute the maximum (see equation 5.5) or the average similarity (see equation 5.6).

To compute the maximum similarity sim_{u_m}, only the tuple $(C_i, D_j) \in S_u$ with the highest similarity value is taken into account. In the second case, sim_{u_w} is the weighted sum of similarities for all tuples (C_i, D_j) (see equation 5.6). The weighting ω can be determined by the count of tuples or by analyzing the ontological structure [79]. The sum of all weights is always 1 ($\sum \omega = 1$). If the similarity of one tuple $(C_i, D_j) \in S_u$ is 1, the weight for this tuple is set to 1 (and all others to 0). This is necessary to keep the equation consistent with definition 5.10. S_u is the set of selected tuples on the intersection level.

The SIM-DL similarity server discussed in chapter 6 supports both similarity functions, however, sim_{u_m} is used by default and applied in the human subject test described in chapter 7.

$$sim_{u_m}(C, D) = max(sim_i(C_i, D_j)); \text{ where } (C_i, D_j) \in S_u \tag{5.5}$$

$$sim_{u_w}(C, D) = \sum_{(C_i, D_j) \in S_u} \omega_{ij} * sim_i(C_i, D_j) \tag{5.6}$$

5.5.5 *Existential Quantification*

In DL, an existential quantification consists of two parts, the role and the filler. The quantification $\exists R.C$ denotes the set of all $a \in \Delta^{\mathcal{I}}$ for which there exists at least one tuple $(a, b) \in R^{\mathcal{I}}$ and $b \in C^{\mathcal{I}}$. Consequently, the similarity between two concepts C and D formed by existential quantification is based on the similarity of the involved roles and fillers. The roles are compared using sim_r. Which similarity function is applied for the fillers depends on their constructors. Following the $\mathcal{ALCHQ}$ normal form (see section 5.3), sim sim_u or sim_i is used in most cases.

Definition 5.13 *The comparison of two concepts C ($\exists R.E$) and D ($\exists S.F$) yields 1, if the similarities $sim_r(R, S)$ and $sim_u(E, F)$ both yield 1 (see figure 16a).*

Given $R, S \in \mathcal{N_R}$ and $C, D \in \mathcal{T}$:
$\simeq_e (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $\simeq_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \wedge \simeq (E^{\mathcal{I}}, F^{\mathcal{I}})$

Definition 5.14 *The comparison of two concepts C ($\exists R.E$) and D ($\exists S.F$) yields a value between 0 and 1, if either the similarity for $sim_r(R, S)$ or $sim_u(E, F)$ yields 1 and the respective other is between 0 and 1; or if both similarity values are between 0 and 1 (see figure 16b).*

Given $R, S \in \mathcal{N_R}$ and $C, D \in \mathcal{T}$:
$\lesseqgtr_e (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $(\simeq_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \wedge \lesseqgtr (E^{\mathcal{I}}, F^{\mathcal{I}})) \vee (\lesseqgtr_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \wedge \simeq (E^{\mathcal{I}}, F^{\mathcal{I}}))$
$\vee (\lesseqgtr_r (S^{\mathcal{I}}, R^{\mathcal{I}}) \wedge \lesseqgtr (E^{\mathcal{I}}, F^{\mathcal{I}}))$

Definition 5.15 *The comparison of two concepts C ($\exists R.E$) and D ($\exists S.F$) yields 0, if at least the similarity of $sim_r(R, S)$ or $sim_u(E, F)$ yields 0 (see figure 16c).*

Given $R, S \in \mathcal{N_R}$ and $C, D \in \mathcal{T}$:
$\not\sim_e (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $\not\sim_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \vee \not\sim (E^{\mathcal{I}}, F^{\mathcal{I}})$

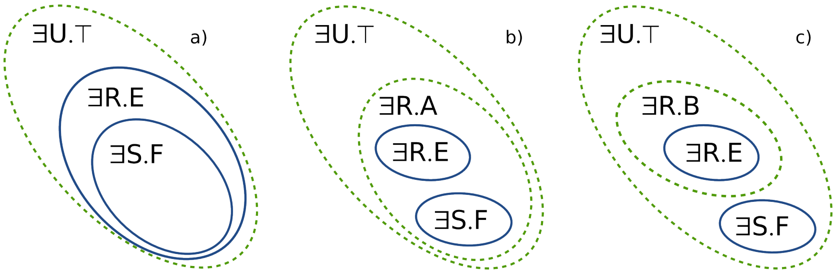

Figure 16: Set-based visualization of similarity between concepts formed by existential quantification: a) $sim_e(C, D) = 1$, b) $sim_e(C, D) =]0, 1[$, and c) $sim_e(C, D) = 0$.

REALIZATION Based on the definitions 5.13, 5.14, and 5.15, sim_e computes the similarity between concepts as the product of role and filler similarity. In addition to this multiplicative approach (see equation 5.7), one can also argue for an averaged sum of role and filler similarity. As discussed in chapter 7, the first approach is chosen for the SIM-DL server as it better approximates human similarity judgment. To avoid a contradiction with definition 5.15 the averaged sum would require a weight of 0 if either the similarity for the roles or fillers is 0.

$$sim_e(C, D) = sim_r(R, S) * sim(E, F) \tag{5.7}$$

5.5.6 *Value Restriction*

In DL, a value restriction consists of two parts, the role and the filler. The restriction $\forall R.C$ denotes the set of all $a \in \Delta^{\mathcal{I}}$, for which every b in a tuple $(a, b) \in R^{\mathcal{I}}$ is an element of $C^{\mathcal{I}}$. Consequently, the similarity between two concepts C and D formed by value restriction is based on the similarity of the involved roles and fillers. The roles are compared using sim_r. Which similarity function is applied to the fillers depends on their constructor. Following the $\mathcal{ALCHQ}$ normal form (see section 5.3), sim_u or sim_i is used in most cases.

Note that in contrast to existential quantification, a value restriction does not state that there is a tuple $(a, b) \in R^{\mathcal{I}}$ for every a, but that if such a tuple exists b is an element of $C^{\mathcal{I}}$.

Definition 5.16 *The comparison of two concepts C ($\forall R.E$) and D ($\forall S.F$) yields 1, if the similarities $sim_r(R, S)$ and $sim_u(E, F)$ both yield 1 (see figure 17a).*

Given $R, S \in \mathcal{N_R}$ and $C, D \in \mathcal{T}$:
$\simeq_f (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $\simeq_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \wedge \simeq (E^{\mathcal{I}}, F^{\mathcal{I}})$

Definition 5.17 *The comparison of two concepts C ($\forall R.E$) and D ($\forall S.F$) yields a value between 0 and 1, if either the similarity for $sim_r(R, S)$ or $sim_u(E, F)$ yields 1 and the respective other is between 0 and 1; or if both similarity values are between 0 and 1 (see figure 17b).*

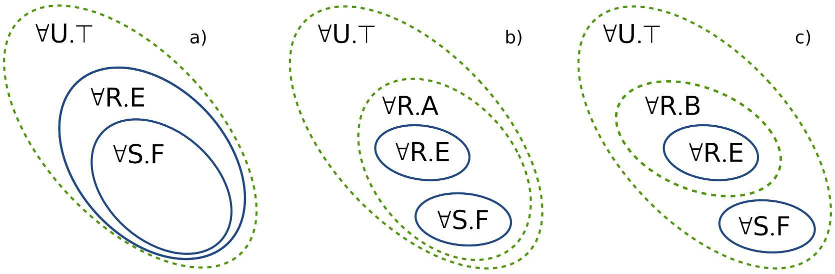

Figure 17: Set-based visualization of similarity between concepts formed by value restriction: a) $sim_f(C,D) = 1$, b) $sim_f(C,D) =]0,1[$, and c) $sim_f(C,D) = 0$.

Given $R, S \in \mathcal{N_R}$ and $C, D \in \mathcal{T}$:
$\lessgtr_f (C^\mathcal{I}, D^\mathcal{I})$ *iff* $(\simeq_r (R^\mathcal{I}, S^\mathcal{I}) \wedge \lessgtr (E^\mathcal{I}, F^\mathcal{I})) \vee (\lessgtr_r (R^\mathcal{I}, S^\mathcal{I}) \wedge \simeq (E^\mathcal{I}, F^\mathcal{I}))$
$\vee (\lessgtr_r (S^\mathcal{I}, R^\mathcal{I}) \wedge \lessgtr (E^\mathcal{I}, F^\mathcal{I}))$

Definition 5.18 *The comparison of two concepts C (∀R.E) and D (∀S.F) yields 0, if at least the similarity of $sim_r(R,S)$ or $sim_u(E,F)$ yields 0 (see figure 17c).*

Given $R, S \in \mathcal{N_R}$ and $C, D \in \mathcal{T}$:
$\nsim_f (C^\mathcal{I}, D^\mathcal{I})$ *iff* $\nsim_r (R^\mathcal{I}, S^\mathcal{I}) \vee \nsim (E^\mathcal{I}, F^\mathcal{I})$

REALIZATION Based on the definitions 5.16, 5.17, and 5.18, sim_f computes similarity between concepts as the product of role and filler similarity (see equation 5.8). As above, the multiplicative approach is taken for sim_f.

$$sim_f(C,D) = sim_r(R,S) * sim(E,F) \tag{5.8}$$

5.5.7 *Qualified Number Restriction*

In DL, a qualified number restriction consists of three parts, the cardinality, the role and the filler. The restriction $(\geq n\, R.C)$ denotes the set of all $a \in \Delta^\mathcal{I}$ for which every b in a tuple $(a,b) \in R^\mathcal{I}$ is an element of $C^\mathcal{I}$, and there are at least (or at most, respectively) n such tuples. Consequently, the similarity between two concepts C and D formed by qualified number restrictions is based on the similarity of the involved roles, fillers, and the cardinality. The roles are compared using sim_r. Which similarity function is applied for the fillers depends on their constructor. The following definitions focus on the min restriction, the max restriction is specified accordingly.

Definition 5.19 *The comparison of two concepts C ($\geq n\, R.E$) and D ($\geq m\, S.F$) yields 1, if the similarities $sim_r(R,S)$ and $sim_u(E,F)$ both yield 1 and $n \leq m$ (see figure 18a).*

Given $R, S \in \mathcal{N_R}$, $C, D \in \mathcal{T}$ and $n, m \in \mathbb{N}_0$:

$\simeq_m (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $\simeq_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \wedge \simeq (E^{\mathcal{I}}, F^{\mathcal{I}}) \wedge n \leq m$

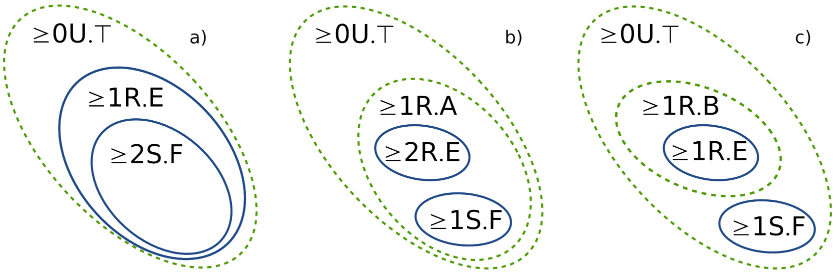

Figure 18: Set-based visualization of similarity between concepts formed by qualified number restriction: a) $sim_m(C, D) = 1$, b) $sim_m(C, D) =]0, 1[$, and c) $sim_m(C, D) = 0$.

Definition 5.20 *The comparison of two concepts C ($\geq n\, R.E$) and D ($\geq m\, S.F$) yields a value between 0 and 1, if the similarity values for $sim_r(R, S)$ and $sim_u(E, F)$ are both between 0 and 1 (or one of them is 1, while the other is]0,1[). If both are 1, $n > m$ (see figure 18b).*

Given $R, S \in \mathcal{N_R}$, $C, D \in \mathcal{T}$ *and* $n, m \in \mathbb{N}_+$:
$\lessgtr_m (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $(\simeq_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \wedge \lessgtr (E^{\mathcal{I}}, F^{\mathcal{I}})) \vee (\lessgtr_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \wedge \simeq (E^{\mathcal{I}}, F^{\mathcal{I}}))$
$\vee (\lessgtr_r (S^{\mathcal{I}}, R^{\mathcal{I}}) \wedge \lessgtr (E^{\mathcal{I}}, F^{\mathcal{I}})) \vee (\simeq_r (S^{\mathcal{I}}, R^{\mathcal{I}}) \wedge \simeq (E^{\mathcal{I}}, F^{\mathcal{I}}) \wedge n > m)$

Definition 5.21 *The comparison of two concepts C ($\geq n\, R.E$) and D ($\geq m\, S.F$) yields 0, if the similarity of $sim_r(R, S)$ or $sim_u(E, F)$ yields 0 or if either n or m are 0 (see figure 18c).*

Given $R, S \in \mathcal{N_R}$, $C, D \in \mathcal{T}$ *and* $n, m \in \mathbb{N}_0$:
$\not\sim_m (C^{\mathcal{I}}, D^{\mathcal{I}})$ *iff* $\not\sim_r (R^{\mathcal{I}}, S^{\mathcal{I}}) \vee \not\sim (E^{\mathcal{I}}, F^{\mathcal{I}}) \vee (n = 0 \oplus m = 0)$

REALIZATION Based on the definitions 5.19, 5.20, and 5.21, sim_m defines similarity between concepts as the product of role, filler, and cardinality similarity. While the roles and fillers are treated the same way as for existential quantifications and value restrictions, cardinalities have to be taken into account. The symbol m is used as abbreviation for min, and max, respectively, indicating that the same equation is applied for both cases. $m_{RS}(total)$ denotes the highest maximum (respectively minimum) cardinality for the roles R or S in the user defined context (see equation 5.8). Similarity between number restrictions therefore depends on their relative distance, where $mRS(total)$ reflects the notion of universe known from statistics. If either n or m is 0, the equation 5.9 does not apply, and $sim_m(C, D) = 0$.

$$sim_m(C, D) = sim_r(R, S) * \left(1 - \frac{\mid m_R(C) - m_S(D) \mid}{m_{RS}(total)}\right) * sim_u(E, F) \tag{5.9}$$

5.5.8 *Inclusion*

While SIM-DL focuses on concept definitions ($\equiv$), inclusion axioms ($\sqsubseteq$) play a crucial role for most DL-based ontologies. In many cases it is not possible to clearly separate a concept from its superconcepts. This is for two reasons. First, the expressivity of the used language may not be sufficient to state the necessary facts. Second, it may be difficult to determine (or agree on) the necessary and sufficient descriptors which distinguish both concepts. These difficulties seem to be typical for both abstract and specific concepts, while they appear less often on the level of basic categories [103, 121, 144, 174]. Nevertheless, it makes sense to distinguish such concepts. Instead of an explicit definition, one can state that a given concept is a specialization [6] of its superconcepts (such as for $D \sqsubseteq C$). Following the description in section 5.3, $D \sqsubseteq C$ is unfolded to $D \equiv C \sqcap D'$. To compare concepts formed by inclusion, both superconcepts the ancestor and the primitiveness (C and D') have to be taken into account. To determine their similarity sim_p is used. With respect to normalization, we assume that the primitiveness corresponds to exactly one descriptor (which is a rough approximation).

To compare the concept *Transportation* to *Supply* (both specializations of $Infrastructure$), sim_p has to be measured between the tuples $(Transportation', Supply')$ and $(Infrastructure, Infrastructure)$. If there is no additional concept defined as *Transportation* and *Supply* infrastructure, the similarity of the first tuple is 0 and therefore the overall similarity yields 0.5. In this example we assume that $Infrastructure$ is a primitive, if otherwise it is replaced by its definition.

5.5.9 *Negation*

$\mathcal{ALCHQ}$ allows for the negation of arbitrary concepts. The following section shows how SIM-DL compares concept descriptions involving negation.

The similarity $sim_p(A, \neg A)$ between a primitive concept and its negation is always 0. A concept and its negation always form an alignable difference (see section 2.1.2); hence, in the presence of $\neg A$, A is compared to its negation and not to other primitives. For a concept $C \equiv A \sqcap B$ and its negation, the similarity $sim(C, \neg C)$ is determined by measuring $sim((A \sqcap B), (\neg A \sqcup \neg B))$. Consequently, similarity between a complex concept (formed by intersection of two primitives) and its negation is always 0. Double negation is handled the same way. To compare C to $\neg D$ ($D \equiv A \sqcap \neg B$), the similarity $sim((A \sqcap B), (\neg A \sqcup B))$ is measured.

In terms of Baader's family ontology, if *Woman* is defined as $Person \sqcap Female$ and *Man* is specified as $Person \sqcap \neg Female$, $sim(Woman, \neg\, Man)$ is measured as follows. First, both concepts are unfolded and rephrased according to the $\mathcal{ALCHQ}$ normal form. In a second step, the similarity $sim((Person \sqcap Female), (\neg Person \sqcup Female))$ is determined by comparing *Person* to $\neg$ *Person* and *Female* to *Female*. The resulting similarity is 0.5. The other way around (and taking asymmetry into account), if $\neg Woman$ is the search concept $sim(\neg Woman,\ Man)$ is 1.

5.5.10 *Context as Universe of Discourse*

Within SIM-DL, context has an impact on both concept selection (see section 5.2) and comparison (e.g., via sim_p or sim_m). All target concepts are subsumers of the context concept, and therefore, share some descriptors – namely those forming the context concept. Following Rodríguez' variability (context) weighting [143], the importance of particular descriptors decreases, if they are part of all concepts selected for comparison. To model this observation, SIM-DL does not take those descriptors of compared concepts into account, which are descriptors of the context concept. If the context concept is defined as *Waterbody*, all target concepts compared to the user's search concept are necessarily waterbodies; hence, *Waterbody* is temporarily removed during the alignment process[5]. As a result, the concepts *River* and *Canal* are less similar within the context of waterbodies than within a context of geographic features in general.

5.5.11 *Conceptual Neighborhood*

Modeling conceptual neighborhoods [22, 27, 37, 38, 45] in description logics is a rarely addressed issue so far. Lutz and Möller [110] investigated how to enable spatial reasoning (with topological relations) using standard description logics. This section shows, how similarity is measured between roles organized within conceptual neighborhoods instead of subsumption hierarchies. While we focus on topological and temporal neighborhoods here (see also figure 3), the proposed measure can be used for any kind of conceptual neighborhood. The only pre-condition is that compared roles are represented within a neighborhood graph.

Definition 5.22 *The comparison of two roles R and S within a conceptual neighborhood yields 1, only if $R = S$.*

Definition 5.23 *The comparison of two roles R and S within a conceptual neighborhood yields a value between 0 and 1, if there exists a path from R to S which is shorter than the longest path within the graph representation of the conceptual neighborhood (and $R \neq S$).*

Definition 5.24 *The comparison of two roles R and S yields 0, if R and S are not within the same conceptual neighborhood or if the shortest path between both roles is at the same time the longest path within the graph representation of the conceptual neighborhood.*

REALIZATION Based on the definitions 5.22, 5.23 and 5.24, similarity between roles (sim_n) is their normalized distance within the graph representation of the conceptual neighborhood. In contrast to sim_r, the normalization is not depth-dependent but the longest path within the neighborhood graph.

5 At the moment, SIM-DL only supports primitives (and their intersections) as context concepts to avoid more complex substraction (difference) operations on the compared concepts (see [20, 162] for further details).

$$sim_n(R,S) = \frac{max\_distance_n - edge\_distance(R,S)}{max\_distance_n} \quad (5.10)$$

5.6 OVERALL SIMILARITY

Within SIM-DL, each similarity function takes care of its normalization using the number of compared tuples or a graph depth. Each similarity function returns a value between 0 and 1 to the function (on a higher level) it was called by (see also chapter 6 and appendix A).

In the following and based on the previous sections, some properties of SIM-DL's notion of similarity are discussed. While the relations $\simeq$, $\lesssim$, and $\nsim$ were used to introduce the formal definitions of particular similarity functions, this section focuses on the overall similarity between compared concepts. For many years, there has been an ongoing discussion on the characteristics of similarity in general and specific measures in particular. This section points out how SIM-DL positions itself within this discussion. We focus on the relation to dissimilarity, strictness, and symmetry leaving reflexivity, transitivity, and the triangle inequality aside.

5.6.1 *Dissimilarity*

One may assume that dissimilarity is exactly the counterpart of similarity: $dis(C,D) = 1 - sim(C,D)$. While this may be true for certain cases, it is not a valid assumption in general [56]. As argued by Tversky [168], Nosofsky [130], and Dubois and Prade[36], both are different views on stimuli comparison. SIM-DL puts much stress on the alignment of descriptors and this alignment process is not reversible. If the task is to find dissimilarities between compared concepts, other tuples might be selected for comparison.

By using the maximum similarity function on union level (sim_{u_m}; see section 5.5.4), one can demonstrate that the assumption $dis(C,D) = 1 - sim(C,D)$ is oversimplified and counter intuitive. Consider the concepts $C \equiv A \sqcap B$ and $D \equiv C \sqcup E$ where A, B, and E are primitives. To measure the similarity $sim(C,D)$, SIM-DL creates the following tuples: (A,A), (A,B), (A,E), (B,A), (B,B), and (B,E). Out of this set, the tuples (A,A) and (B,B) are chosen for further computation and finally, $sim(C,D)$ returns 1. Consequently, the resulting dissimilarity $dis(C,D)$ should be 0. This is true, if one still applies the maximum similarity function. Instead, when searching for dissimilarities between compared concepts one would rather use a minimum similarity function and thus take E into account ($dis(C,D) > 0$).

5.6.2 *Strictness*

Strictness is often referred to as an important property of similarity [161]. Formally, strictness states that the maximum similarity value is only assigned to equal stimuli (concepts): $sim(C,D) = 1\ iff\ C \equiv D$. This is related to the minimality property which claims that two different stimuli are less (or equally) similar[6] than the stimulus is to itself : $sim(C, C) \leq sim(C, D)$ [4, 56].

In SIM-DL, the similarity value 1 is interpreted as 'equal or not distinguishable (within a given context)'. This is for two reasons, co-occurrence between primitives and asymmetry. The comparison of two primitives yields 1, if they cannot be differentiated, i.e., if they always appear jointly within concept definitions (see section 5.5.1). As SIM-DL focuses on information retrieval, a target concept satisfies the users needs ($sim(C_s, C_t) = 1$) if it is a subclass of the search concept (see 5.4). Consequently, similarity is not strict.

5.6.3 *Symmetry*

Symmetry is one of the most controversial properties of similarity. While several theories from computer science argue that similarity is essentially a symmetric relation [108], research from cognitive science favors asymmetric similarity measures [98, 123, 130, 168]. As argued in the previous sections, while SIM-DL supports symmetry if requested by the user[7], it is defined as an asymmetric measure by default. From Tversky's point of view, one may argue that allowing both approaches is nothing more than indecision. However, the understanding of symmetry underlying SIM-DL is driven by Nosofsky's notion of a biased measure [130]. Asymmetry is not a characteristic of similarity as such, but of the process of measuring similarity. This process is driven (biased) by a certain task - namely information retrieval. Whether the comparison of two concepts involves asymmetry or not depends on the application area and task (and therefore the alignment process), but not on the measure as such.

5.7 Comparison to Related Similarity Measures

The following section compares the SIM-DL measure to related similarity theories and non-standard inference techniques such as the *Least Common Subsumer* (lcs).

5.7.1 *Class Level versus Instance Level Similarity*

Answering the question whether a particular measure compares individuals or concepts does not necessarily answer the question whether similarity is a class or instance level relationship. If the compared concepts are feature lists such as used in MDSM [143], each individual exactly consists of the features specified within the concept definition. Consequently, $sim(C, D) = sim(c, d)$ for all $c \in C$ and $d \in D$. If the compared concepts are prototypes or models (such as in SIM-DL), the previous assumption cannot be made. In the first case, radial categories [103] are a classical counterexample. In the second case, one may think of concept descriptions involving unions, value or number restrictions. In Baader's family ontology, *Parent* is specified as $Mother \sqcup Father$. While the similarity $sim(Parent, Parent) = 1$, $sim(f, m) \neq 1$ if f is an

6 In the literature, minimality is defined for dissimilarity: $dis(C, D) \geq dis(C, C)$.
7 via the similarity server and Protégé plug-in discussed in chapter 6.

instance of *Father*, while m is an instance of *Mother*. SIM-DL treats similarity as a class level relationship, which at first view seems to contradict the definition proposed by Lin [108].

> The semantic similarity between two classes C and C' is not about the classes themselves. When we say "rivers and ditches are similar", we are not comparing the set of rivers with the set of ditches. Instead, we are comparing a generic river and a generic ditch. Therefore, we define $sim(C, C')$ to be the similarity between x and x' if all we know about x and x' is that $x \in C$ and $x' \in C'$. [108, p. 301]

With respect to SIM-DL, when we say "the concepts River and Ditch are similar", we argue that the characteristics (i.e., set restrictions) used for categorization of particular geographic features are similar. Concepts are formal specifications for the boundaries restricting the membership of individuals to a particular category (see also [54, 115, 153]). Consequently, if the restrictions posed on the categories $River^{\mathcal{I}}$ and $Ditch^{\mathcal{I}}$ are similar, the members of these categories should be similar[8] (see also [173] and [84]). Based on this assumption, we state that a prototypical river should be similar to a prototypical ditch, without assuming that all rivers and ditches are equally similar (nor that SIM-DL measures similarity between a prototypical river and ditch). For example, let us assume that rivers are defined as watercourses with at least one spring as origin and one waterbody as destination. Still, one may think of a specific application ontology mostly populated by rivers connected to more than two hydrographic features. These rivers may be atypical and not very similar to (prototypical) rivers within other ontologies; nevertheless, they satisfy the criteria for category membership. In the absence of comparable individuals (and this is the kind of scenario SIM-DL is made for), "we define $sim(C, C')$ to be the similarity between x and x' if all we know about x and x' is that $x \in C$ and $x' \in C'$."[108, p. 301]

5.7.2 *DL-Based Measures*

SIM-DL works exclusively on the terminological level, while some related approaches also take individuals into account [32, 33]. To make this point clear, consider the following example. The concepts C ($C \equiv A \sqcap B \sqcap \exists R.A$) and D ($D \equiv A \sqcap \neg B \sqcap \exists R.A$) shall be compared. An instance matching algorithm would compute the similarity $sim(C, D)$ based on their extensions, i.e., as the ratio of common and distinct individuals . The resulting similarity depends on the extensions of A, B, and $\exists R.A$. Similarity changes, if the extensions of the compared concepts change. Such a measure is especially useful for tasks such as clustering. In contrast, SIM-DL computes the similarity $sim(C, D)$ based on their definitions (i.e., intensional). Simplified, $sim(C, D)$ yields 0.67[9] independently of the extensions of C and D. In terms of Baader's family

8 Note that, while the notion of category is reduced to a set theoretic view here, categories have structure. For instance, categories have a graded structure, while sets are unstructured.

9 $sim(C, D) = \frac{sim(A,A)+sim(\exists R.A,\exists R.A)+sim(B,\neg B)}{3} = \frac{2}{3}$

ontology, if A is *Person*, B *Female*, and R *hasChild*, the instance based similarity between C and D could differ from a case were A is *Animal*. In SIM-DL, similarity does not change because it relies on structural comparisons.

Still, one should not argue that one of both approaches is more appropriate without defining the application area. If all instances of the compared concepts are known and defined within the same ontology, an instance matching approach may be more accurate than SIM-DL. In contrast, SIM-DL can reason about concept similarity in the absence of individuals. Additionally, while both approaches measure similarity, they answer slightly different questions. The first approach [32, 33] assumes that concepts are the more similar, the more individuals they share (at execution time). In contrast, SIM-DL answers the question of whether potential individuals are restricted by the same membership constraints.

In addition, most related measures only compare such quantifications and restrictions which are based on the same role, while SIM-DL also measures similarity between roles.

5.7.3 *Least Common Subsumer*

Besides classical subsumption reasoning, non-standard inference gained interest within the last years [7, 102]. This includes unification and matching, but especially techniques to determine the least common subsumer (*lcs*) and most specific concept [102]. Further approaches find commonalities among selected concepts or individuals [113, 114] and approximate concept descriptions [20]. This section shows how SIM-DL relates to the least common subsumer. Formally, the *lcs* can be defined as follows:

Definition 5.25 *Given a description logic $\mathcal{L}$, and a set of concepts $C_1, ..., C_n$, a particular concept D is the least common subsumer with respect to $C_1, ..., C_n$ iff it satisfies the following conditions:*

(a) $C_i \sqsubseteq D$ for all $C_{1,...,n}$
(b) All concepts D' satisfying $C_i \sqsubseteq D'$ (for all $C_{1,...,n}$) also satisfy $D \sqsubseteq D'$, i.e., D is the least $\mathcal{L}$ concept satisfying (a) and unique.

First of all, the context concept C_c used within SIM-DL (see section 5.2) is comparable to the least common subsumer of all target concepts ($\{C_t | C_t \sqsubseteq C_c\}$). The only difference between the *lcs* and C_c is that the context concept is defined by the user while its subconcepts are inferred. In contrast, the *lcs* is defined the opposite way, - the concepts are known while the least common subsumer is inferred. In both cases, the concept does not need to be a named concept within the examined ontology.

To measure similarity, one could compute the least common subsumer of compared concepts and apply a network-based measure afterwards to determine their distance. This approach is taken by Rodríguez and Egenhofer for MDSM [143], and also to define role similarity within SIM-DL. For primitives, SIM-DL uses Jaccard's coefficient, because computing the least common subsumer does not deliver meaningful results here (in fact, this is true for all expressive description logics using disjunction; see [8] for approximations of *lcs*). Given two primitives A and B, $lcs(A, B)$ is simply $A \sqcup B$, which does not add any new information. Finally, while classical similarity theories map concepts to a real number, the *lcs* maps concepts to their subsumer (which is again a concept).

Another alternative how the *lcs* can be used to define similarity was proposed by Möller et al.[102, 126]. In a first step, the user defines a query by selecting some exemplary individuals from the ABox. Next, the concepts of these individuals are retrieved[10]. Out of these concepts the least common subsumer is inferred as so-called retrieval concept. Finally, the result of the user's query is the set of all individuals that instantiate the retrieval concept. Compared to SIM-DL, such a measure relies on individuals and does not deliver a ranking (through similarity values). It is especially useful, if the task is to find related individuals by pointing out some desired examples.

5.7.4 *Feature and Network Based Measures*

Roughly speaking, SIM-DL is a combination of feature (sim_p) and network-based (sim_r) approaches to similarity. All further similarity functions (such as sim_i or sim_e) are combinations of these measures. This section points out why pure feature or network-based approaches are not sufficient to compare DL concepts.

Classical feature-based measures only compare values (instead of role-filler pairs) within unstructured lists of (untyped) features. Rodríguez and Egenhofer [143] extended this view by proposing typed features (parts, functions, and attributes). Still, functions, parts, and attributes remain on the level of unary predicates (e.g., play, green,...) without a further definition. Additionally, feature-based measures distinguish between common and distinct features, without the possibility to define partial matches[11] (see section 2.1.1). A description logics-based measure built upon feature similarity was proposed by Borgida et al. [18] for the (less expressive) $\mathcal{A}$ language.

As role-filler pairs cannot be represented, one may try to group them as single features such as *overlap_Waterbody*. For several reasons such an approach points to the wrong direction. First, this raises the question of what a feature is [18]. In terms of the classical feature theory and as argued above, the features chosen on a concept level hold for all instances. This is clearly not the case for DL concepts such as $(\geq 2\, overlap\ Waterbody)$. Additionally, it is not clear whether there should be a single feature for DL expression, or several (e.g., $min\_2\_overlap\_Waterbody$). Second, comparing such features would result in counter intuitive similarity values. Independently of whether $min\_1\_overlap\_Waterbody$ or $min\_2\_overlap\_River$ is chosen as target, the resulting similarity would always be 0. Finally, DL primitives could not be matched directly and hence their comparison would return 0.

10 This step can also be done by computing the most specific concept (*msc*) for each individual [102]

11 An extended version of MDSM allowing for partial matches by introducing thematic roles was presented by Janowicz [77].

In contrast, SIM-DL uses feature matching indirectly to compute the co-occurrence of primitives within complex concepts.

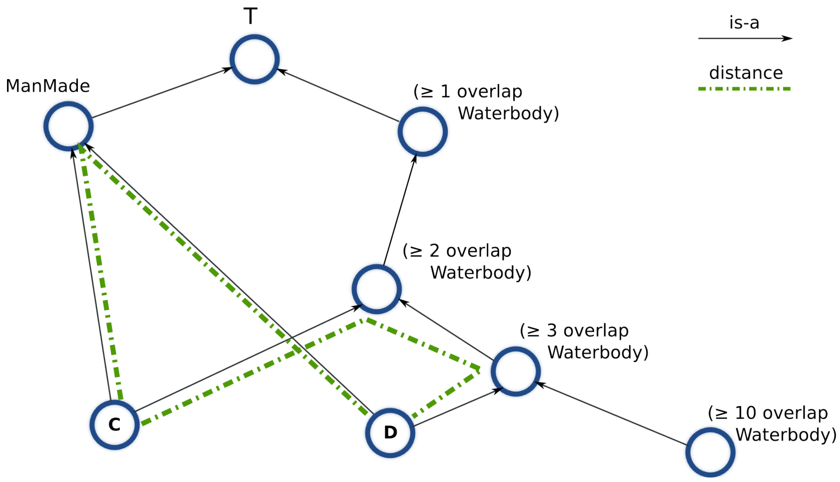

Figure 19: Similarity as inverse distance within a is-a hierarchy. There are two possible paths for sim(C,D). In addition, the distance between the concepts formed by number restrictions would be equal.

Network-based measures compute similarity using a concept graph. Roughly speaking, similarity is defined as the shortest path between the compared concepts normalized by the maximum distance within this graph. Starting with the measure proposed by Rada et al. [137], dozens of network-based similarity theories propose slightly different approaches taking the depth of $lub(C, D)$, edge density or different edge weights into account [173]. Additionally, while most measures focus on is-a hierarchies, further models also consider partonomic relations (see also [143]). Most network-based theories are applicable for comparing concepts within single inheritance structures, but fail in the case of multiple inheritance. If one tries to incorporate additional roles or expressive language constructors, their impact (and weight) within the graph becomes unclear (see figure 19). For instance, a network-based measure would return equal similarity values between $(\geq\ 10\, overlap\ Waterbody)$ and $(\geq\ 3\, overlap\ Waterbody)$, and between $(\geq\ 3\, overlap\ Waterbody)$ and $(\geq\ 2\, overlap\ Waterbody)$. To avoid such difficulties, SIM-DL uses a weighted network measure only to compute similarity between roles. These roles are atomic and represented within a single inheritance hierarchy.

5.7.5 *Measures for Fuzzy Sets*

Similarity measurement plays an important role within research on fuzzy sets. A current overview of theories and applications was given by Cross and Sudkamp [30] as well as by Chen and colleagues [26]. This section argues why SIM-DL does not incorporate these theories.

Most similarity theories introduced for fuzzy sets are in fact generalizations of measures defined for classic (crisp) sets obtained by employing fuzzy (membership) operations [36]. This includes feature-based[12], geometric, network, and information theoretic measures. In contrast to classical set theory, the degree of membership, i.e., whether a particular entity is within a set or not, is not a boolean but real valued function (see definition 5.26). The more an entity belongs to a fuzzy set, the higher is its degree of membership.

Definition 5.26 *Given the universe of discourse $U = \{u_1, ..., u_n\}$, a fuzzy set C (over the universe) is defined by its membership function $\mu_C : U \to [0, 1]$. A membership value $\mu_C(u_i)$ describes to which degree u_i is a member of C.*

Within SIM-DL, fuzzy set-based measures are not used for two reasons. First, these measures rely on similarity theories for crisp sets. What was argued for feature and network-based approaches to similarity (and about the difference between instance and concept level) before, is therefore also true for fuzzy set theories. Second, the definition of elementary operators differs from those used within description logics (and classical set theory, respectively). For instance, the subsets are specified as follows:

Definition 5.27 *Given two sets C and D and their membership functions μ_C and μ_D, C is a subsets of D ($C \subseteq D$) iff $\forall u \in U$, $\mu_C(u) \leq \mu_D(u)$.*

5.7.6 *Formal Concept Analysis-based Similarity Measures*

Formal concept analysis (FCA) is a framework based on Lattice theory intended for analyzing, structuring, and visualizing data [23, 47, 135]. Given a formal context defined as triple $\mathbb{K} := (E, A, I)$ over a set of entities E, a set of attributes A, and an indication relationship I ($I \subseteq E \times A$), a formal concept is specified as follows[13]:

Definition 5.28 *A FCA concept X is defined as a tuple (E_i, A_i) where:*

- $E_i \subseteq E$
- $A_i \subseteq A$
- $E_i' = A_i$; *where* $E_i' = \{a \in A | eIa\, \forall e \in E_i\}$
- $A_i' = E_i$; *where* $A_i' = \{e \in E | eIa\, \forall a \in A_i\}$

The set E_i is also called the *extent* of the concept X, A_i its *intent*. In contrast to classical ontologies, these both cannot be separated, i.e., while most ontologies define concepts via their intention, FCA concepts always consists of entities and the attributes that these entities share. Consequently, the SIM-DL theory cannot be applied to FCA concepts nor can FCA-based similarity measures be used for DL based ontologies (at least for those without defined (named) individuals). An inter-concept similarity measure for Formal Concept Analysis has been introduced by Formica [42].

12 Often referred to as set-theoretic measures within fuzzy set research.
13 If eIa, for an $e \in E$ and $a \in A$, then the entity e possesses the attribute a.

5.8 summary

Based on the framework specified in section 2.2, the SIM-DL theory was defined as an asymmetric and context-aware similarity measure, which compares DL concepts by computing the overlap between their descriptors. As these descriptors are concepts themselves, the recursive process terminates when only primitive concepts and roles are left. Their similarity is determined by network (sim_r and sim_n) and feature-based (sim_p) measures [137, 143].

6 SIM-DL SIMILARITY SERVER AND PLUG-IN

This chapter introduces the SIM-DL similarity server and Protégé plug-in. The server is developed as an implementation of the SIM-DL theory, while the plug-in acts as interface for the Protégé ontology editor. Both have been developed within the SimCat project, and are available as (Java-based) free and open source software at sourceforge.net[1]. The chapter refers to the version beta2.2 of the server and plug-in. New releases focus on fixing bugs and adding support for more expressive description logics.

The chapter discusses the server architecture, required extensions to the DIG description logic interface (DIG 1.1) [15, 35], and the SIM-DL Protégé plug-in. For each part of the server architecture, a reference to the theoretical foundations described in chapter 5 is given. A user manual for the server and plug-in is also available at the SimCat project webpage.

6.1 ARCHITECTURE

The SIM-DL server is based on an embedded Jetty HTTP server[2] to handle XML and is listening on port 8085 (TCP) per default. Incoming requests via XML-over-HTTP are processed by a request handler which interprets DIG operations and starts the similarity and reasoning engines. The reasoner implements a tableaux algorithm to determine TBox subsumption based on ABox satisfiability [6, 72, 74, 124]. The similarity engine is based on the presented SIM-DL framework and theory. Both components implement their own normalization and blocking methods [72, 124]. Caching and lazy unfolding are implemented to reduce execution time and memory usage [72].

The activity diagram depicted in figure 20 shows the steps involved in processing similarity queries. The diagram starts where the request handler has received the queried ontology, the search concept, and target concept or context concept, respectively. If the query defines a target concept explicitly, the search and target concepts are passed to a canonization function (see section 5.3). If the query contains the search concept and context concept (and an optional threshold), the built-in subsumption reasoner has to determine the target concepts (see section 5.2) beforehand. In the next step, a subsumption hierarchy is built based on all involved concepts. This taxonomy is necessary to create the alignment matrix (see section 5.4) in the next step, and is also used for some of the similarity functions, such as sim_p (see section 5.5). SIM-DL supports symmetric and asymmetric similarity. Depending on the user's query, different alignment matrices have to be created, and different normalization factors are computed (to return values between 0 and 1 for every similarity function). If either the search or target

1 http://sim-dl.sourceforge.net/downloads/
2 http://jetty.mortbay.org/

concept contains logical disjunction, the user can decide between two similarity modes: average and maximum similarity.

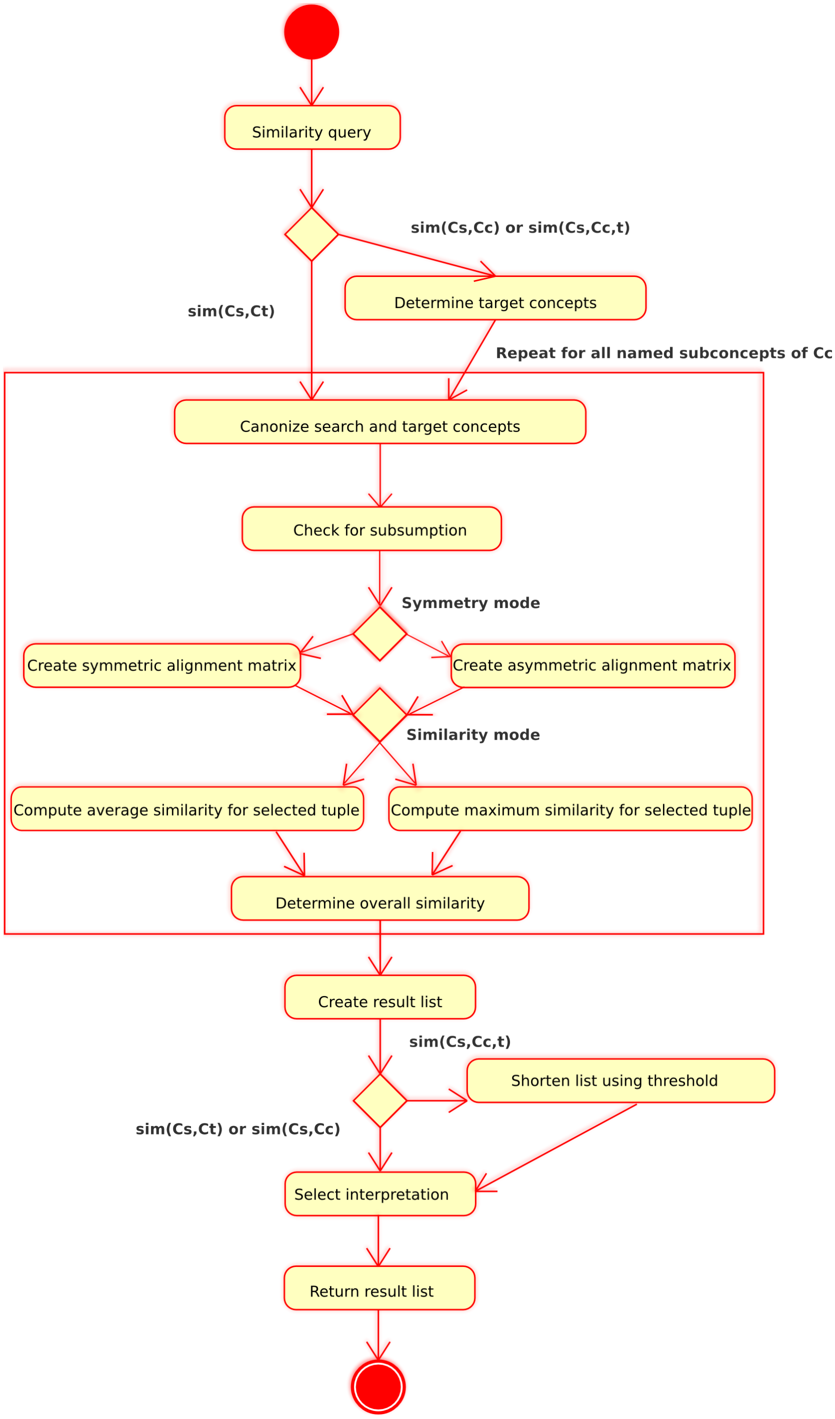

Figure 20: UML activity diagram showing how the SIM-DL server processes similarity queries.

In the first case, all concepts combined by the union constructor are compared and influence the overall similarity. In the second case, only the tuple with the highest similarity is taken for further computation (see section 5.5.4). Depending on the constructors and concepts used within the definitions of the search and target concept, this step also involves all other similarity functions discussed in chapter 5. Finally, the overall similarity is computed and the whole process starts again until all target concepts have been compared to the search concept.

This leads to an ordered list of tuples containing each target concept and its similarity value with respect to the search concept. If the query was defined using only one particular target concept, the list contains one entry. Two further steps are necessary before the results can be presented to the user. First, if a threshold was defined, the list has to be truncated. Second, the ordered list has to be transformed to a *similarity ranking*. This step is called interpretation. One possible interpretation was discussed in chapter 4, namely font size scaling. Another interpretation is a classification of the results (where the number of categories is specified by the user).

The SIM-DL server supports additional kinds of contexts which influence similarity; these are left out here for reasons of readability. They are covered in section 8.4 together with more details on the interpretation of similarity values.

6.2 DIG-EXTENSION

A short introduction to the DIG interface was given in section 3.2.3. The interface has to be extended to enable similarity measurement between concepts as specified in SIM-DL. First, the *Ask* syntax has to be extended by a similarity query which defines the search concept (C_s) and the context concept (C_c), or target concept (C_t), respectively. Table 3 shows the supported queries as well as the similarity extension (for the case where a context concept is chosen).

A *ccsimilarity* query requires the following elements: The symmetry and the similarity modes have to be defined. The default values are *asymmetric* and *maxSimilarity*, respectively. This is the typical setting used for information retrieval. If required, these modes can be changed using the keywords *symmetric* and *avgSimilarity*. Each query has to specify the threshold and the number of categories. The default value for both elements is 1. The threshold ranges from 1 to 100, which corresponds to similarity values between 0.01 and 1. This is to avoid that concepts are returned which have no overlap with the search concept. An arbitrary number of categories can be requested. The actual number of categories returned to the user depends on the number of compared target concepts and their similarity values, i.e., concepts with the same similarity value are necessarily in the same category. Table 3 illustrates the situation when the search concept is a named concept within the ontology, while the context concept is defined within the *Ask* query (e.g., as intersection of two named concepts). This is not mandatory, both concepts can be either named concepts or defined in the query. In contrast, the target concept has to exist beforehand.

Table 4 shows the supported DIG response operators and the extension which permits responses to similarity queries. The result of

Table 3: Supported Ask syntax and similarity (ask) extensions.

Request Category	Ask Syntax
Satisfiability	`<satisfiable>C</satisfiable>`
Concept Hierarchy	`<parents>C</parents>`
	`<children>C</children>`
	`<ancestors>C</ancestors>`
	`<descendants>C</descendants>`
	`<equivalents>C</equivalents>`
Similarity Queries	`<ccsimilarity>`
	`    <symmetryMode name="asymmetric|symmetric">`
	`    <similarityMode name="maxSimilarity|avgSimilarity">`
	`    <threshold value=Integer>`
	`    <category value=Integer>`
	`    <searchConcept>`
	`        <catom name=CS>`
	`    </searchConcept>`
	`    <contextConcept>`
	`        <and>`
	`            <catom name=CC1>`
	`            ...`
	`            <catom name=CCN>`
	`        </and>`
	`    </contextConcept>`
	`</ccsimilarity>`
	`<ctsimilarity>`
	`...`
	`</ctsimilarity>`

a similarity query is a set of concepts grouped into categories. Each category has an indexing number; the first category contains the most similar target concepts, the second category consists of the second best, and so on. Each category contains at least one target concept and all of those are named concepts. Each concept within a category has a similarity value and a font scaling indicator associated. These scalings can be transformed into font sizes within a particular application, such as the SIM-DL Protégé plug-in. The font size itself depends on the used font, whose selection is up to the application (or the user).

One could argue that the interpretation of the results should be up to the application contacting the similarity server. In a such case, the server would only deliver similarity values. Both approaches have their advantages and disadvantages. We decided to implement these elements within the server for two reasons. First, both interpretations are useful within several scenarios, as demonstrated using the gazetteer Web interface and the SIM-DL Protégé plug-in. Second, this should reduce the barrier in creating third party applications which use the similarity server within their workflow. Additionally, the server always delivers the uninterpreted similarity values. Therefore, one can implement application-dependent interpretations (or representation) layers.

Table 4: Supported response syntax and similarity (response) extensions.

Response Category	Response Syntax	Request Category
Boolean	<true/>	Satisfiability
	<false/>	
Concept Set	<conceptSet>	Concept Hierarchy
	<synonyms>S11...S1N</synonyms>	
	<synonyms>SM1...SMN</synonyms>	
	</conceptSet>	
Similarity Ranking	<conceptSet>	*Similarity Query*
	<category index=1>	
	<catom name=S1>	
	<simValue>s1</simValue>	
	<fontSize>f1</fontSize>	
	</catom>	
	...	
	<catom name=SN>	
	<simValue>sN</simValue>	
	<fontSize>fN</fontSize>	
	</catom>	
	</category>	
	...	
	<category index=M>	
	...	
	</category>	
	</conceptSet>	

6.3 SIM-DL Protégé Plug-in

Graphical user interfaces are an important aspect for communication with reasoning services. This includes classical reasoning such as subsumption reasoning, but also non-standard inference such as similarity reasoning. Today's de facto standard frontend for description logics-based reasoning is the open source Protégé ontology editor. Protégé is built upon an extensible architecture which provides the possibility to add further functionality via plug-ins. The Protégé OWL plug-in is the most popular extension; it enables users to create, explore, and modify OWL ontologies, and supports OWL-Lite, OWL-DL, and OWL-Full [96]. In addition, it provides DIG-based access to description logics reasoners such as Pellet [152] and FaCT++ [166]. The combination of description logics, reasoner support, and the Protégé editor as graphical frontend was a prerequisite for establishing OWL as standard for creating semantic Web applications. Besides editing ontologies, Protégé was also extended to support visualization and information retrieval.

A comparable combination of tools and theory is necessary to initiate the spread of description logics-based similarity measurement. The Protégé OWL API includes several extension points to implement OWL specific plug-ins. To provide a graphical frontend for querying the SIM-DL similarity server, we developed a plug-in based on Protégé's OWL capabilities. The functionality to view and explore the ontology used for similarity measurement is necessary. This is already provided by the Protégé OWL extension, and is reused within the SIM-DL plug-in. Due to the architecture of the similarity server the SIM-DL plug-in has to support DIG as communication interface. We integrated the DIG implementation provided by Protégé OWL and added the SIM-DL specific DIG elements (see table 3 and 4).

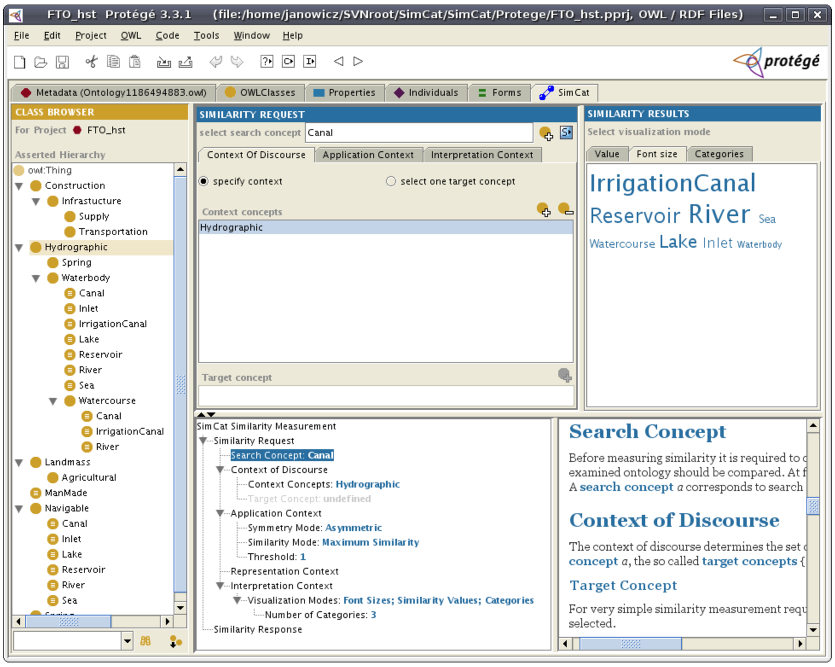

Figure 21: Selection of search concept and context concept (called context of discourse here ; see section 8.4 for details) using the SIM-DL Protégé plug-in.

The plug-in is depicted in figure 21; it already contains the extended context model which is introduced in section 8.4. The plug-in is structured into the following five frames. The leftmost frame displays the concept hierarchy of the used ontology. The two frames in the middle are used to create the similarity query. The upper frame consists of three tabs to formulate the query, while the bottom frame displays a navigable tree view on the current settings (specified in the frame above). Each element of the tree is a link which displays (in the bottom right frame) a short introduction and some background information about the kind of context or similarity mode, respectively. Finally, the top-right frame displays the results. In figure 21, it contains three tabs, as three kinds of interpretations were selected for the query.

In the first tab (of the query frame), the user selects a search concept and either specifies a context of discourse (the context concept) or a particular target concept. In the following tab, named *Application Context*, the user can choose the symmetry and similarity mode, and define a threshold. This corresponds to the mode specifications for the extended DIG *Ask* syntax. Using the *Interpretation Context* tab, the three standard outputs of the SIM-DL server can be selected. The results can be either presented as a descending list of similarity values (between

0 and 1), using font size scaling, or split up into categories. In the last case, the user can define the number of categories.

Future versions of the plug-in will focus on the integration of the so-called *representation context* (see section 8.4). This requires research on how to create the necessary rules (see [92] for details) and is therefore left for future work. The plug-in is developed for ontology engineers familiar with the used vocabulary and description logics. The integration into an end-user centric interface, such as the gazetteer web interface, was discussed in chapter 4.

7 APPLICATION AND EVALUATION

This chapter introduces two experiments [81, 86] performed with human participants to verify whether the similarity rankings obtained using the SIM-DL theory introduced in chapter 5 positively correlates with human similarity rankings. SIM-DL is intended to measure similarity between computational representations of concepts. The motivation is to improve the accessibility of tasks such as information retrieval and organization for human users. This can only be achieved if there is a high correlation between the similarity rankings calculated by SIM-DL and human similarity judgments. The SIM-DL measurement process was developed based on findings from cognitive science. It takes aspects such as asymmetry, alignment, and context into account which are known to play an important role for human similarity ratings. SIM-DL tries to approximate aspects from the human process of reasoning about similarity to achieve meaningful results. Nevertheless, it is a computational theory for description logics rather than a framework for understanding cognitive processes. Consequently, we neither claim that SIM-DL models the process of human similarity judgments nor that humans represent concepts in any kind of logic-based form.

Figure 22 illustrates the relation between a similarity reasoning service such as the SIM-DL server and human reasoning about similarity. The box at the top represents the cognitive process (marked as dotted line) of deriving similarity judgments. Without discussing the relationship between representation and human cognition in detail [50, 118], up to now no direct mapping to computational representations is possible. Similarity theories developed in cognitive science model (i.e., approximate) this process by partitioning it into observable units. The effect of each unit is studied by changing its settings, while all other units remain stable[1]. Such units include Context, Alignment, Asymmetry, and the Max-Effect [123]. Each of them is depicted as a box on the dotted process line to indicate that they are fragments of the whole process. Most theories from cognitive science focus on the explanation of human similarity reasoning rather than the development of executable services[2]. In contrast, information science is interested in computational representations to provide a basis for executable theories. While these theories approximate cognitive theories, their goal is not necessarily explanatory. Instead, they adopt elements that can be computed with appropriate resources. From this point of view, computational models form a subset of theories established in cognitive science. Typical application areas include human computer interaction and information retrieval.

The box at the bottom of figure 22 represents concrete similarity reasoning services such as the SIM-DL similarity server. These services implement the computational theories as standalone applications or as

1 Or by studying patients with lesions.
2 For some exceptions, see SME and MAC/FAC [40, 49].

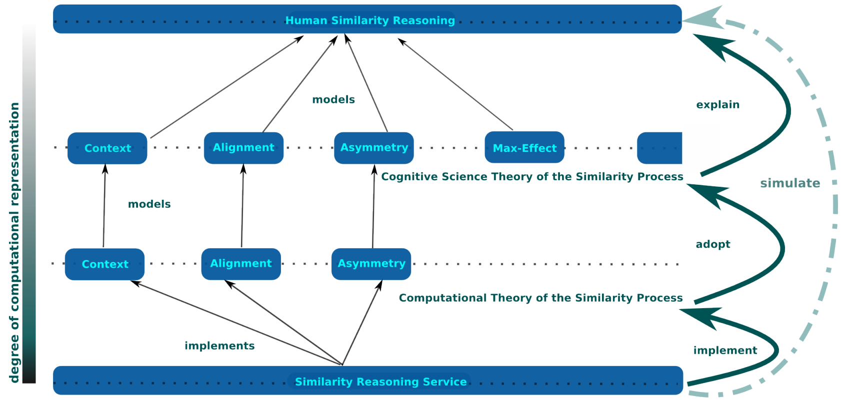

Figure 22: From human similarity reasoning to similarity services such as the SIM-DL similarity server (from Janowicz et al. [86]).

parts of a knowledge infrastructure such as the ConceptVISTA[3] toolbox. The motivation for developing similarity-aware applications is to simulate human similarity judgment, thus making tasks such as information retrieval more accessible to the user. It is important to note that not the cognitive process is simulated, but the final similarity ranking, i.e., the reasoning results. The dashed arrow indicates that there is no direct link between the similarity service and human similarity judgments. Computational similarity ratings depend on how compared entities and concepts are represented and which units (parts) of the human similarity process are modeled within the implemented computational theories.

The term *cognitively plausible* is used, if the similarity rankings produced using SIM-DL correlate with human similarity rankings. In contrast, *cognitively adequate* similarity measurement would require a comparison of the underlying processes and is out of scope for this work.

7.1 ROLES AND FILLERS

This section presents a human participants test, examining whether subjects prefer a multiplicative or additive approach to role-filler similarity.

7.1.1 *Motivation*

The comparison of concepts formed by existential quantifications, value restrictions, or quantified number restrictions requires a combined similarity measure for roles and their fillers (see also section 5.5). While it is possible to define in which cases the resulting similarity should be 1,]0,1[, or 0, from an information scientific point of view it is not possible to decide whether a multiplicative or additive approach should be taken. Both fulfill the formal requirements specified for SIM-DL, but produce different similarity values in case of overlapping concept descriptions.

3 http://www.geovista.psu.edu/ConceptVISTA/

First, in the multiplicative case, the comparison of $\forall inside.Canal$ to $\forall inside.Ocean$[4] yields a similarity value of 0.3. In case of an additive approach, i.e., the arithmetic mean, the similarity would be 0.65. A third possibility would be a weighted average with self adjusting weights. The results for such an approach would vary between 0.3 and 1 depending on the weights. To find out which approach should be implemented within the SIM-DL similarity server, a Web-based human participants test has been carried out [81].

7.1.2 *Test Setting*

First, three possible similarity functions have been implemented, the multiplicative approach (see equation 7.1), the arithmetic mean (see equation 7.2)[5], and the auto-weighted mean (see equation 7.3). sim_r corresponds to the role similarity, while sim_f is the similarity for the fillers (and should not be confused with the similarity for value restrictions introduced in chapter 5).

$$sim_{mult}(R.E, S.F) = sim_r(R, S) * sim_f(E, F) \tag{7.1}$$

$$sim_{am}(R.E, S.F) = \frac{sim_r(R, S) + sim_f(E, F)}{2} \tag{7.2}$$

sim_{wam} is the weighted average of the similarity (sim_r) derived by comparing the roles R to S, and the similarity obtained by measuring the similarity (sim_f) between the fillers E and F. The role and filler weightings (ω_r and ω_f) reflect the relative importance of sim_r and sim_f within sim_{wam} and are defined in terms of the absolute difference between role and filler similarity. If the inter-role and inter-filler similarities are close together, both have a similar impact on sim_{wam}. Otherwise, the lower similarity value gets a higher weighting [81][80]. This weighting function is chosen here, because the resulting similarity values are higher than those obtained using the multiplicative approach, but lower than those computed using the unweighted mean (if $sim_r \neq sim_f$).

$$sim_{wam}(R.E, S.F) = \frac{\omega_r * sim_r(R, S) + \omega_f * sim_f(E, F)}{\omega_r + \omega_f} \tag{7.3}$$

$$where \begin{cases} sim_r \geq sim_f & \omega_r = 1 - |sim_r - sim_f|;\ \omega_f = 1 + |sim_r - sim_f| \\ sim_f > sim_r & \omega_f = 1 - |sim_r - sim_f|;\ \omega_r = 1 + |sim_r - sim_f| \end{cases}$$

A Web-based questionnaire has been developed which compares human similarity judgments to those obtained by the three similarity functions. The questionnaire consists of five tabs (see figures 23-27) and has been prepared in German. Starting from the first tab, the participants move on until the last tab is reached. It is not possible to move backwards, i.e., previous results can neither be seen nor changed.

4 If we assume that the similarity between *Canal* and *Ocean* is 0.3 within a given context.
5 Note that if $sim_r(R, S)$ or $sim_f(E, F) = 0$, $sim(R.E, S.F)$ is set to 0 (see also 5.5.5).

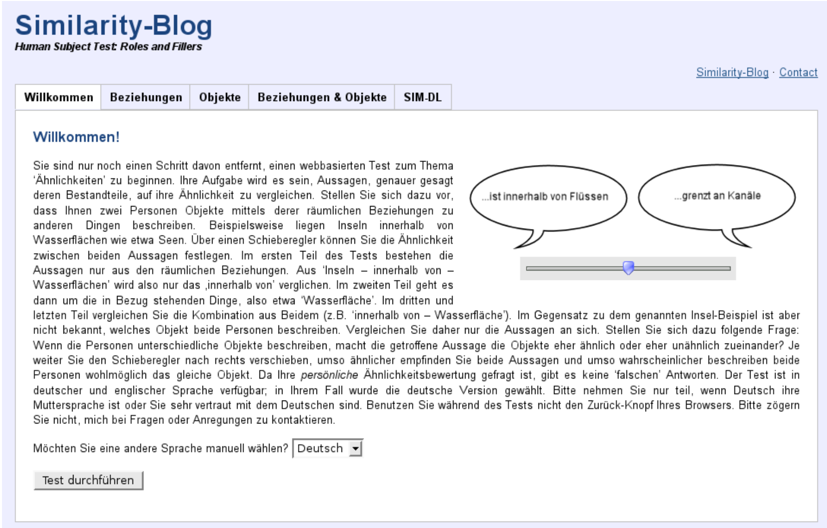

Figure 23: The welcome tab introducing the test setting and motivation.

While similarity and categorization are language dependent, this plays a minor role within the presented test setting. German has been chosen, because it is the native language of all participants and helps to avoid misunderstandings.

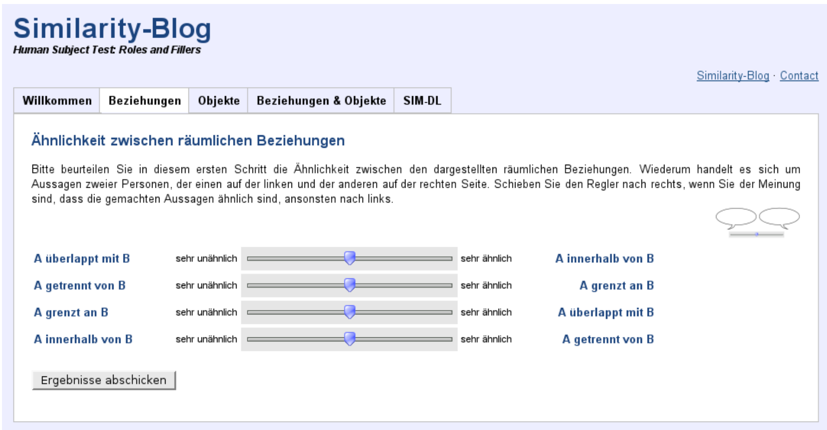

Figure 24: The second tab displaying topological relations.

The first tab is titled *Welcome* (ger.: Willkommen) and introduces the task (see figure 23). The participant is asked to rate how similar two statements are. These statements describe topological relations such as *islands are within waterbodies (e.g., lakes)*. These statements are not directly presented to the user, but separated into three steps (each on a single tab). In the first step, the user is comparing topological relations as such, e.g., *inside* to *meets*. In the second step, the related objects are compared for similarity (e.g., *river* to *canal*). In the third step, combined statements of the type *inside river* are compared.

To ensure that the user only focuses on the relations in the first step, the related objects (in fact subject and object) are replaced by placeholders. The statement *island - inside - waterbody* is replaced by *A inside B* (see figure 24). The second step only displays the object names without the subject or relation of the statement, i.e., *river* or *canal* (see figure 25). In the third step, relation - object pairs such as *A inside waterway* are displayed to the participant (see figure 26).

Figure 25: The third tab displaying objects (geographic feature types).

After completing the third step, the (combined) similarity judgments of the participant are compared to the results obtained using the similarity functions. These functions take the relation and object (role and filler) similarities from the first two steps to compute the similarities for the last tab (called SIM-DL). As those steps are performed by the participant, no computational representation or reasoning is required. This allows to compare the three SIM-DL functions to the combined human judgments from the third step. To do so, it has to be ensured that the relations and objects rated in the previous steps are presented in the correct combination to the participant.

The *relations* (ger.: Beziehungen) tab displays four pairs of topological relations [22, 27, 37, 38] (e.g., *inside - overlap*) in a randomized order. The possible relations are: *inside* (ger.: innerhalb von), *overlaps* (ger.: überlappt mit), *meets* (ger.: grenzt an), and *disjoint* (ger.: getrennt von). The *objects* (ger.: Objekte) tab displays four pairs of geographic feature types (e.g., *lake - waterbody*) in a randomized order. The possible types are: *canal* (ger.: Kanal), *river* (ger.: Fluss), *lake* (ger.: See), *waterbody* (ger.: Wasserfläche), *watercourse* (ger.: Wasserlauf), and *waterway* (ger.: Wasserstraße). Each relation and object occurs at most once on the right and once on the left side. The pairs established in the relations tab and objects tab are used for the combined statements in the relations-objects tab. For instance, if *A inside B* is compared to *A overlaps B* and *river* to *canal*, then in the following step, *A inside river* is compared to *A overlaps canal*. The problem of naming topological relations was discussed by Riedeman [140], while the influence of scale was discussed by Lautenschütz et al. [106].

To make the comparisons in the test less abstract and artificial for the participants, they were asked to imagine a discussion between two

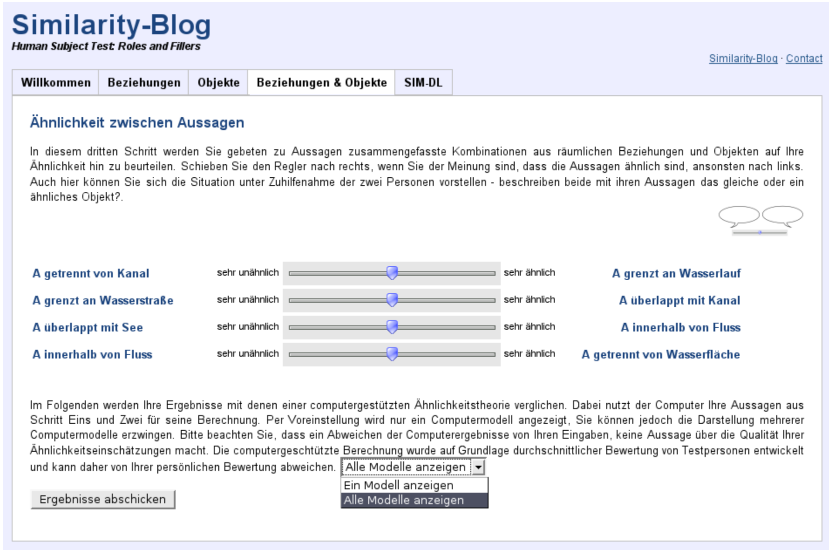

Figure 26: The fourth tab combining relations and objects to statements.

persons. The fictitious persons both describe an object via its topological relation to other objects. Based on their statements, the participant is asked to rate whether both actually describe the same object or not. This has to be done by adjusting a slider between these statements. The more the slider is pushed from the left statement to the right statement, the more similar both statements are and the more probably they talk about the same object (see figure 23). The starting position of the slider is always in the middle, which corresponds to a similarity value of 0.5. The left most position is called *very dissimilar* (ger.: sehr unähnlich) and translates to a similarity value of 0. The right most position is named *very similar* (ger.: sehr ähnlich) and maps to the value 1. This corresponds to the findings discussed in section 5.6.1 and by Medin and colleagues [123]; similarity does not range from dissimilar to equal.

The slider can be moved using the mouse or the left and the right arrow key on the keyboard. The position of the slider can be changed until the tab is finalized by pressing on the *submit results* button (ger.: Ergebnisse abschicken). Font size, positioning, and slider size are adjusted to screen resolutions of 1024x768 (XGA) and 1240x1024 (SXGA). According to Webhits[6], these resolutions are used by more than 80% of the Web users. While the size of the slider bar can differ from screen to screen, it should be large enough to allow for a precise positioning of the slider arrow. To ensure that each user only participates once, the IP-address is saved together with a time stamp. The problem of dynamic IPs is not an issue here, as the test only runs for a short period of time. The results of each participant (see figure 27)[7] are stored in a database for further evaluation.

6 http://www.webhits.de/deutsch/index.shtml?webstats.html (visited 09/2007)
7 The depicted test run was performed for demonstration purpose and was not recorded for further processing.

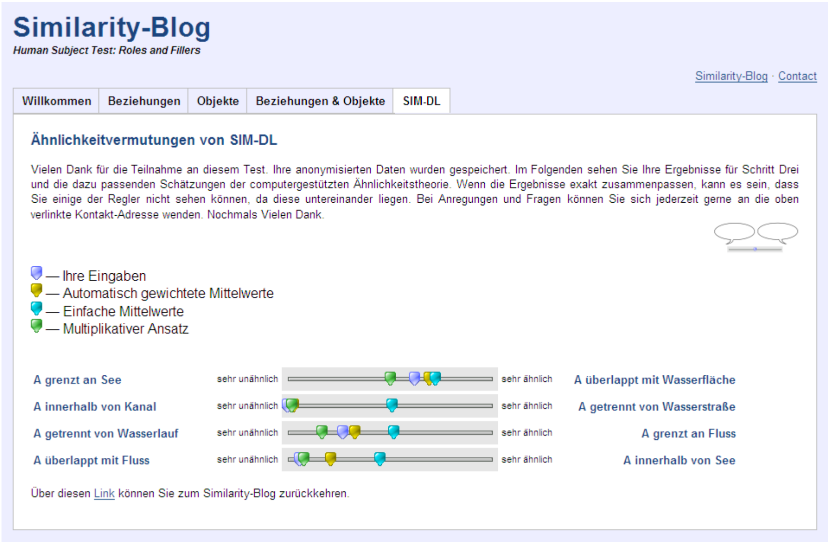

Figure 27: Result tab comparing human estimates to those from SIM-DL.

7.1.3 *Results*

Out of 84 (relation-object) similarity estimations from 21 participants, 80 were taken for further computation[8]. Each human estimation was compared to the results from the three machine-based calculations (using the participant's relation and object estimations from the previous steps) by computing the absolute deviation. The results were grouped into 10 classes ranging from an absolute deviation of 0 up to 100. As the low deviation classes (0-10 and 11-20) were of special interest, they were further divided into three classes ranging from 0-5, 6-11 and 12-17. The results are displayed in table 5 and visualized in figure 28.

Table 5: Absolute deviation of the three similarity functions compared to human estimations. The numbers represent frequencies.

Class	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
sim_{am}	5	11	5	5	24	13	5	5	6	1
sim_{wam}	33	26	5	7	4	2	2	1	0	0
sim_{mult}	41	15	7	8	2	5	1	1	0	0
Class	0-5	6-11	12-17							
sim_{am}	2	3	5							
sim_{wam}	16	20	19							
sim_{mult}	26	15	11							

Compared to the arithmetic mean ($\overline{x}$) of the absolute deviations and the average absolute deviation[9] (AD) for sim_{am} (45.21, 16.77), the re-

8 The participants' gender and age were not collected for this Web-based test.
9 Which, in this case, is the mean deviation from the arithmetic mean of the absolute deviations: $\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$

sults for sim_{wam} (17.85, 12.57) and sim_{mult} (17.16, 14.34) were close. To reveal the differences between sim_{wam} and sim_{mult}, modus and median were computed. The results are presented in table 6. It should also be mentioned that in contrast to sim_{wam} and sim_{mult}, the results from sim_{am} always overestimate, i.e., the deviation is positive.

Table 6: Measures of statistical dispersion for the absolute deviations from human estimations.

	Mean	AD	Median	Mode
sim_{am}	45.21	16.77	45.75	43
sim_{wam}	17.85	12.57	12	12
sim_{mult}	17.16	14.34	10	0

7.1.4 *Discussion*

As depicted in figure 28, the arithmetic mean similarity function sim_{am} differs clearly from the participants' estimations. While the absolute deviation is high in general, it is especially apparent in cases such as the comparison of *inside canal* to *disjoint watercourse*. One participant rated the similarity of *inside-disjoint* to be 0, while the similarity of *canal-watercourse* was rated 0.82. The participant assigned a similarity value of 0 in the relation-object step, while sim_{am} proposes a similarity of 0.41 (see also the second row in figure 27 as example). Using sim_{wam} and sim_{mult} yields a similarity of 0.07 or 0, respectively. Analyzing these results, one could argue that roles and fillers do not have the same impact on similarity estimations and that weights are necessary (which is proposed in the related literature, see chapter 3). The automated weighting function sim_{wam} which assigns the higher weight to the lower similarity value returns more accurate results. Taking sim_{mult} into account changes the picture, because this multiplicative approach is unweighted but accurate.

A detailed analysis of the 80 similarity estimations reveals that in several cases the results for the relation-object step were lower than the similarities for the relations and objects. Such behavior cannot be reproduced using a weighting function. This is not surprising, as it indicates that human similarity judgments are based on causal connections. If there is no similarity between the relations, the similarity of the related objects does not play any role (and vice versa)[10]. The median and mode (see table 6) for sim_{mult} and sim_{wam} support this view. The difference between median and mode for sim_{mult}, is caused by the 0-similarity estimations. In most cases participants judged the relation-object pair 0, if either the relation or object similarity was 0. In some cases the participant's estimations were slightly above 0. This can be explained by the slider used in the Web interface. The slider has to be moved from its initial, middle postion to the left end of the slider bar. The value 0 is returned, when the head and not the body of the arrow reaches the left end (see figure 27). Hence, these estimations can

10 Therefore, one cannot argue that the role is more important than the filler.

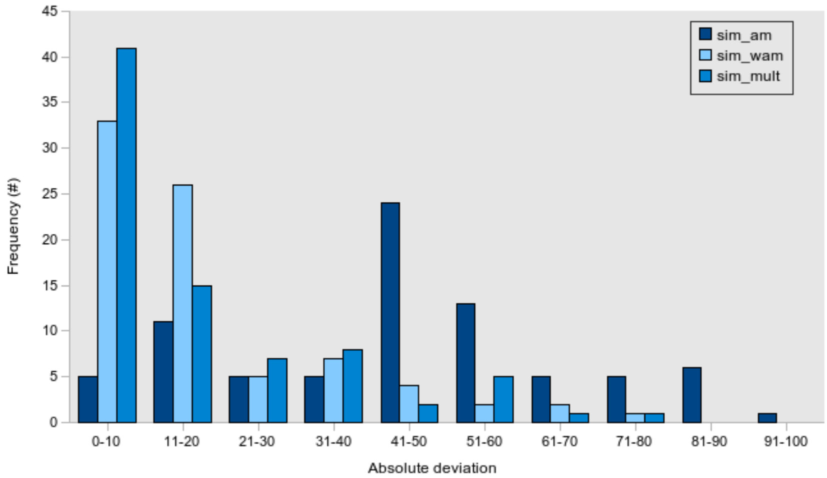

Figure 28: Histogram of the absolute deviation classes.

be either explained by a weakness of the test design or by participants which did not pay enough attention while adjusting the slider.

The next interesting aspect is the arithmetic mean and average absolute deviation for sim_{mult} and sim_{wam}. Compared to mode and median, the mean (absolute deviation) is higher than expected[11]. To find an explanation, the data was analyzed for human estimations which differ clearly (absolute deviation $> 2 * \overline{x}$) from the calculations using sim_{mult} and sim_{wam}. The resulting data set contains two sorts of estimations.

The first sort consists of some data without common characteristics, and is therefore considered as the usual divergence of human judgments or misunderstandings (with respect to the task). An example could be the comparison of *meets waterway* to *inside canal*. A participant rated the similarity *meets-inside* 0.77 and the similarity *waterway-canal* 0.78. While sim_{mult} and sim_{wam} return 0.6 and 0.77, respectively, the participant rated the combined pair 0. The participant's relation rating is high compared to other participants, which may explain the unusual combined estimation.

The second sort of estimations contains estimations which seem to involve spatial inference. For instance, a participant rated the similarity between *disjoint* and *inside* to be 0.06 and the similarity between *lake* and *river* 0.5. The combined similarity for *disjoint lake* and *inside river* was judged 0.76. In contrast, sim_{mult} and sim_{wam} return 0.03 and 0.18, respectively. This surprising kind of judgments may be interpreted as follows. To explain the idea of similarity estimations to the participants, they were told that comparing relation-object pairs could be imagined as rating how probable it is that two people (describing a certain situation in different words) actually talk about the same situation. This explanation may be a reason why some of the participants' similarity estimations were inconsistent and neither captured by the multiplicative approach nor the weighted average. Participants may

11 One has to keep in mind that the robustness of these averages differs with respect to outliers [19].

have assumed that if a described object is inside a lake, it is disjoint from a river. In the above case, the statements to be compared are *A inside lake* and *A disjoint river*.

If we replace *A* with *island*, the participant may have thought of a situation as depicted in figure 29. Asked whether two persons saying *it is inside a lake* respectively *it is disjoint from a river* refer to the same object, i.e., island, this may be the case. Removing such records from the set of considered estimations reduced the arithmetic mean of the absolute deviations and the average absolute deviation. However, it is not clear in which cases and which kind of inference (or preferred model, respectively) was used by individual participants, and therefore, a quantification of this effect is not possible.

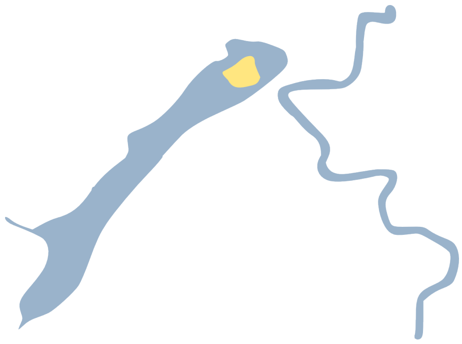

Figure 29: An island which is inside a lake and disjoint from a river.

After evaluating the results, the multiplicative approach sim_{mult} was chosen for SIM-DL as it better approximates human estimations in this test (especially with respect to the classes with a very low absolute deviation). One has to keep in mind that the test was only performed on a small set of topological relations with crisp borders. The situation may be different for other relations and objects (geographic feature types). This leads back to the problem of representation and modelling discussed in the beginning (see also figure 22). Further investigations are necessary to develop more accurate test designs and exclude such cases which involve additional inference.

7.2 SIM-DL AND HUMAN SIMILARITY JUDGMENTS

The following human participants test was performed using sim_{mult} as part of SIM-DL. This section describes the motivation and the test setting. The results are presented and discussed in detail.

7.2.1 *Motivation*

The roles and fillers test was designed to select one of several possible realizations for role-filler similarities within the SIM-DL framework. It does not answer the question whether SIM-DL is *cognitively plausible*, but was an intermediate step to complete the computational similarity theory. The next step was to verify that SIM-DL meets the specified requirements, i.e., whether the ranking calculated by the SIM-DL theory introduced in chapter 5 correlate with human similarity rankings.

Very few tests have been carried out so far concerning the usage of similarity measurement in practice. As the presented similarity theory is related to the Matching Distance Similarity Measure, our test setting was designed according to the more extensive human participants test carried out by Rodríguez [142]. While both tests cannot be compared due to the different representation languages, the new test should be evaluated using the same methods. This includes Spearman's rank correlation coefficient ρ [19], Kendall's rank correlation coefficient τ [19], and Kendall's coefficient of concordance W [90]. This new test has been designed in a way to overcome two difficulties in interpreting the results of Rodríguez' test with respect to SIM-DL.

First, the SIM-DL test is restricted to one domain (hydrographic feature types; see chapter 4) with stepwise decreasing similarity values spanning over the whole interval from maximal to minimal similarity. Rodríguez' test includes comparisons such as *Lake* to *Pond* and *River*, but also to *Desert*, *City*, *Mountain*, and *Bridge* (see [142, Appendix Part E]; the test proposed here is therefore designed according to Part C). While the similarity decreases from *(Lake, Pond)* to *(Lake, River)*, it is difficult to think about *(Lake, Bridge)* or *(Lake, Desert)* in terms of similarity. If *Lake* and *Bridge*, or *Lake* and *Desert* do not share common features, how should users assign ranks to them? Rodríguez' test results (see [142, p. 91]) indicate that subjects rated *Lake* to be more similar to *Bridge* than to *Desert*. We would argue that comparing such entities (or types) involves analogical reasoning [40, 41, 49, 156, 157], semantic relatedness [68], or the ad-hoc construction of scenarios. A bridge may span over a lake so that both are *related*, but not similar. From a similarity point of view, one would rather assume deserts to be more similar. Deserts and lakes, for example, are both not man-made and habitats (see also [142, p. 84]). In terms of analogy, bridges and lakes may be thought of as transportation infrastructure.

This leads to the second difficulty. While it is hard to exclude several kinds of human reasoning in favor of similarity, the SIM-DL test requires a clear mapping from the representation of stimuli in the test setting to the representation used for the similarity computation. For instance, the plain text descriptions handed out to the participants in Rodríguez' test contained descriptions as follows [142, p. 160]:

- **Stadium:** large often unroofed structure in which athletic events are held.
- **Lake:** body of (usually fresh) water surrounded by land.
- **Bridge:** structure erected over a depression or obstacle to carry traffic or some facility such as a pipeline.

First, these definitions contain terms such as *often*, *usually*, and *such as*. It is not clear how they are mapped to the computational representation used by MDSM [143]. The term *often unroofed* is mapped to *covered/uncovered* (see [142, p. 60, 114] and [143]). This does not necessarily reflect the character of *often*, indicating that there are more uncovered than covered stadiums. The Backus-Naur form (BNF) for

MDSM [143] does also not allow to express *usually* or *such as*[12]. Second, the descriptions handed out to the participants were a subset of the definitions used by MDSM. In case of *Stadium*, the latter includes attributes such as *owner_type*, parts such as *dressing_room*, and functions such as *recreate*. Summing up, it is difficult to determine to which degree, the computational representations relate to the participants' representations (or preferred models, respectively). This leads to the question whether human and machine-based similarity rankings are comparable and strictly speaking violates our notion of *cognitively plausible* defined for the SIM-DL test.

7.2.2 *Test Setting*

Due to the fact that the participants were all native German speakers, the test was in German, too. In the following, those parts that are necessary for understanding the test design are translated into English. 28 participants were recruited for the human participants test, 16 males and 12 females. The mean age of the 28 participants was 27.3 with a range from 22 to 31 years. The mean female age was 26.4 and the mean age of the males was 27.8. The questionnaire[13] was distributed randomly among the participants [127].

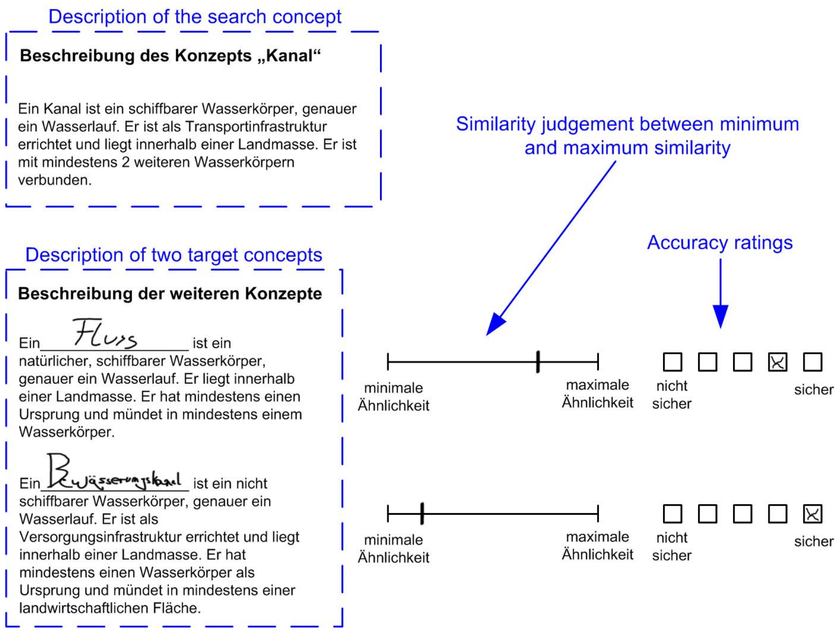

Figure 30: Part of the questionnaire, showing the search concept *Canal* (ger.: Kanal) and two of the six target concepts, *River* (ger.: Fluss) and *Irrigation Canal* (ger.: Bewässerungskanal). The depicted questionnaire was filled out for demonstration purpose and is not part of the human participants test (from Janowicz et al. [86]).

The first step for every participant was to read the introduction, con-

12 Which is an enumeration from on ontological point of view.
13 The questionnaire is available at http://sim-dl.sourceforge.net/downloads/.

sisting of a motivation for the test and instructions on how to complete it. According to Harrison [66], written instructions are preferred by participants over spoken instructions. Next, every participant was asked to read the concept descriptions of the given feature types: the named search concept *Canal* and a set of anonymous target concepts (figure 30). Every participant was requested to assess the similarity between the description of the search concept and every description of the target concepts by placing a mark between a line ranging from minimum to maximum similarity. Additionally, the participants made a statement how confident they felt when placing the mark using a discrete scale with five classes from *not sure* (ger.: nicht sicher) to *sure* (ger.: sicher). We assume that a continuous scale for assessing the concept similarity is reasonable due to the provided granularity[14] which is not required for the confidence assessments. The range for the continuous scale went from minimum similarity (ger.: minimale Ähnlichkeit) to maximum similarity (ger.: maximale Ähnlichkeit). The reason for omitting the names of the target concepts was to ensure that the similarity judgments only depend on the concept descriptions and are not biased by the participants individual conceptualizations.

In the final step, the participants were asked to assign a given list of (concept) names to the anonymous descriptions. This step was introduced to check whether the presented concept descriptions correspond to the participants' conceptualization. Moreover, wrong assignments of the names are a strong hint that the test was filled in randomly, and thus is useless for the evaluation; this check was considered necessary as there was no financial compensation for the participants' effort.

To elucidate the concept descriptions used for the participants test and their representation within the test ontology, three concepts are described in detail[15]. The conceptualizations were carefully derived from the following thesauri and typing schemata: the Alexandria Digital Library Feature Type Thesaurus[16], the Getty Thesaurus of Geographic Names[17], the feature type ontology provided by GeoNames[18], and appendix A from the International Hydrographic Organization Transfer Standard for Digital Hydrographic Data S-57 [75]. Besides the search concept *Canal* (ger.: Kanal), the six target concepts were, *River* (ger.: Fluss), *Irrigation Canal* (ger.: Bewässerungskanal), *Reservoir* (ger.: Stausee), *Lake* (ger.: See), *Ocean* (ger.: Ozean), and *Offshore Platform* (ger.: Förderplattform). The concepts *Canal*, *River*, and *Irrigation Canal* were described as follows[19]:

> A canal is a navigable body of water, namely a watercourse. It is constructed as transportation-infrastructure, and is inside of a landmass. It is connected to at least two other bodies of water.

---

14 For example, to allow statements such as "the similarity between *Canal* and concept A is almost equal to the similarity between *Canal* and concept B, but the former seems to be a bit higher".

15 For further information please download the questionnaire and the feature type ontology at http://sim-dl.sourceforge.net/downloads/.

16 http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/index.htm

17 http://www.getty.edu/research/conducting_research/vocabularies/tgn/

18 http://www.geonames.org/ontology/

19 The underlined text is missing in the printed version of the questionnaire (see figure 30).

> A river is a natural, navigable body of water, namely a watercourse. It is inside of a landmass. It has at least one spring as origin and at least one body of water as destination.
> An irrigation canal is a non-navigable body of water, namely a watercourse. It is constructed as supply-infrastructure and is inside of a landmass. It has at least one body of water as origin and at least one agricultural area as destination.

The ontological counterparts were specified using the $\mathcal{ALCHQ}$ description logic and the Protégé ontology editor.

$$
\begin{aligned}
Canal \sqsubseteq\ & Waterbody \sqcap Watercourse \sqcap Navigable \sqcap (\exists inside.Landmass) \\
& \sqcap (\exists constructedAs.Transportation) \\
& \sqcap (\geq 2\, connectedTo.Waterbody)
\end{aligned}
$$

$$
\begin{aligned}
River \sqsubseteq\ & Waterbody \sqcap Watercourse \sqcap Navigable \sqcap (\exists inside.Landmass) \\
& \sqcap (\neg ManMade) \sqcap (\geq 1\, hasOrigin.Spring) \\
& \sqcap (\geq 1\, hasDestination.Waterbody)
\end{aligned}
$$

$$
\begin{aligned}
IrrigationCanal \sqsubseteq\ & Waterbody \sqcap Watercourse \sqcap (\neg Navigable) \\
& \sqcap (\exists constructedAs.Supply) \sqcap (\exists inside.Landmass) \\
& \sqcap (\geq 1\, hasDestination.AgriculturalFeature) \\
& \sqcap (\geq 1\, hasOrigin.Waterbody)
\end{aligned}
$$

$$Watercourse \sqsubseteq Waterbody \sqcap (\geq 2\, connectedTo.GeographicFeature)$$

$$Waterbody \sqsubseteq HydrographicFeature$$

$$ManMade \sqsubseteq (\exists constructedAs.\top)$$

$$Transportation \sqsubseteq Infrastructure$$

$$Supply \sqsubseteq Infrastructure$$

$$connectedTo \sqsubseteq hasOrigin$$

$$connectedTo \sqsubseteq hasDestination$$

In a first attempt (and pre-test), we tried to use a Controlled Natural Language (CNL) to map OWL code to plain English [28, 150]. This results in complex and artificial sentences which cannot be presented to participants without background in logics. Additionally, the resulting descriptions may force the participants to follow the list-like structure instead of the content during comparison. The concept *IrrigationCanal* would be described as follows:

```
Iff X is an IrrigationCanal,
   then X
    is a Waterbody,
   and X
    has some Waterbody as origin,
    and has some AgriculturalFeature as destination,
    and ...
```

To achieve comprehensible natural language concept descriptions that correspond to the ontology, each part of the concept definition was mapped to a sentence. No additional information (such as *usually* or *often*) was given within the plain text description. The selected sentences were repeated in the same way and order for all concepts. The only exception was the use of *and* to connect sentences. To point out subsumption relations between concepts, constructs such as *It is constructed as transportation-infrastructure* or *[...] body of water, namely a watercourse* were used. This mapping does not exactly preserve the meaning of logical constructors such as existential quantification. Strictly speaking, $\exists inside.Landmass$ allows to be inside of more than just one landmass, which is not reflected in the plain text description *It is inside of a landmass*. To fix this kind of inaccuracies, more detailed DL specifications would be required for the feature types, which would also result in more complex plain text descriptions. This would conflict with a major tenet of the test, namely to keep the cognitive load for participants low. The definition of the seven concepts has been restricted to not more than seven descriptors (e.g., *Navigable*). A higher number of concepts would tire the participants. A higher number of descriptors would result in a longer plain text description and probably force participants to focus on certain parts of the descriptions, leaving others aside.

Some information was left implicit, to ensure that participants do not only perform a syntax-based matching. This includes three cases. First, *natural* (ger.: natürlich) was used as opposite of *constructedAs*. Only artifacts can be constructed[20]. The second case involves basic topological reasoning. *hasOrigin* and *hasDestination* are subroles of *connectedTo*, which was not emphasized in the plain text descriptions. If a watercourse has a body of water as destination, both are necessarily connected. Finally, some basic numerical reasoning was involved when comparing descriptions such as *at least two* and *at least one*.

Note that as the participants were only comparing hydrographic features, *HydrographicFeature* was set as context concept (C_c) within SIM-DL and does not influence the measured similarity. Consequently, only the so-called primitiveness of concepts [72] such as *Waterbody* was considered; see chapter 5 for more details.

7.2.3 *Results*

Out of the 28 questionnaires, 26 were taken for further processing. First, it was checked whether the concept names were properly assigned to the descriptions. All 26 questionnaires satisfy this requirement, however, several participants made updates (changed the names) while performing the test. Next, the similarity values and confidence assessments were transformed to values and weights, respectively, between 0 and 1. Each confidence box corresponds to a weighting step of 0.2. The first box was transformed to 0.2, the second to 0.4, and so on. Table 7 shows the absolute similarity values obtained using the SIM-DL similarity server, the arithmetic mean of the human similarity judgments, and the weighted mean using the confidence assessments.

20 Leaving God as constructor aside.

Table 7: Mean (absolute) similarity judgments between *Canal* and the target concepts.

Concept	Fluss (River)	Bew.kanal (Irr. Canal)	Stausee (Reservoir)	See (Lake)	Ozean (Ocean)	Fdr.plattform (Off. Platform)
Similarity to Canal						
SIM-DL server	0.75	0.67	0.58	0.5	0.38	0.08
Arithm. mean	0.7	0.53	0.59	0.43	0.33	0.14
Weighted mean	0.72	0.55	0.6	0.43	0.32	0.13

In a next step, the absolute similarity values from each questionnaire were transformed to ordinal scale, i.e., into a descending similarity ranking. The most similar concept (with respect to *Canal*) was ranked 6, while the least similar got the rank 1. If two or more concepts had the same absolute similarity values, a mean rank (tie) was chosen (e.g., 4.5). The weights have no influence on the ranking position. Figure 31 shows the resulting box plot for the 26 questionnaires. It depicts the lowest non-outlier ranking, the lower quartile (25%), the median, upper quartile (75%), and highest non-outlier ranking per target concept. The stars and dots represent mild and extreme outliers. *River*, *Reservoir*, *Lake*, and *Ocean* have a comparable interquartile range, while the box plot for the *Offshore Platform* is collapsed. In contrast, the *Irrigation Canal* box plot show a high distribution among test subjects.

Table 8: Median and mode similarity ranks for the target concepts based on the test results.

	Fluss (River)	Bew.kanal (Irr. Canal)	Stausee (Reservoir)	See (Lake)	Ozean (Ocean)	Fdr.plattform (Off. Platform)
N	26	26	26	26	26	26
Median	5.00	5.00	4.00	3.00	2.00	1.00
Mode	5.00[a]	6.00	4.00	3.00	2.00	1.00
frequency (#)						
6th rank	**12**	**8**	6	0	0	0
5th rank	**12**	6	4	1	1	0
4.5th rank[b]	1	-	1	-	-	-
4th rank	1	2	**12**	10	1	0
3rd rank	0	3	1	**14**	7	2
2nd rank	0	3	2	1	**16**	3
1st rank	0	4	0	0	1	**21**

a: Multiple modes exist (5 and 6). The smallest value is shown.

b: This rank is caused by the normalized ranking process.

As depicted in table 8, the individual ranking data from each questionnaire was used to compute the median and mode for each target concept. In both cases, the resulting order corresponds to the computed similarity ranking except that *River* and *Irrigation Canal* share the same rank. In terms of frequencies, this means that the majority of test subjects has chosen the same rank as SIM-DL for *Reservoir*, *Lake*, *Ocean*, and *Offshore Platform*. In case of *River*, the same number of participants has chosen the 6^{th} and 5^{th} rank (12 times), while SIM-DL ranks *River* as most similar concept to *Canal* (6^{th} rank). The remaining two participants selected the 4^{th} rank. While the median for *Irrigation Canal* corresponds to the computed 5^{th} rank, the mode is 6. This is caused by the high dispersion for this concept. The human rankings range from the first (4 times) up to the sixth rank (8 times).

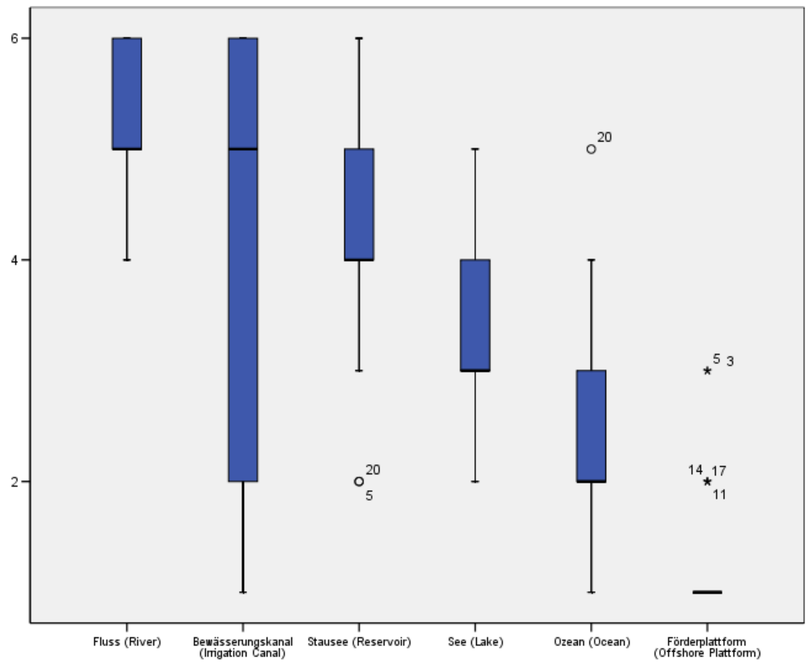

Figure 31: Box plots showing the human similarity rankings and their dispersion. The numbered symbols represent (extreme) outliers (from Janowicz et al. [86]).

A correlation analysis between the median human similarity ranking and the ranking computed by SIM-DL yields $r_s = 0.986$ ($p < 0.01$) using Spearman's ρ. As depicted in figure 31, the data is not normally distributed, i.e., skewed. In addition, we cannot assume equi-distance between the ranks. Hence, the correlation was also determined using Kendall's τ and yields 0.966 ($p < 0.01$). To measure the consensus among participants with respect to the chosen rank, Kendall's coefficient of concordance W was used. To determine whether an obtained W value is significant, chi-square was computed for given degrees of freedom and compared to significance tables for probability. The analysis (taking the ties from the ranking process into account) yields a value of 0.632 for W with a *Chisq(5)* of 82.1. If we hypothesize that the participants' ranks are associated, this corresponds to a proability of p < 0.01 that we accept the hypothesis while it is false. Consequently, and with respect to the number of participants, the results are considered significant[21].

21 We assume a significance level of $\alpha = 0.05$ for this test; see [19, p. 114].

7.2.4 *Discussion*

The test shows a strong and significant correlation between human similarity rankings and those obtained using the SIM-DL similarity server. It also proves that the participants' rankings are associated. Based on our previous definition, the computed similarity judgments can be called *cognitively plausible* (with respect to this test).

The correspondence between the absolute similarity values is difficult to interpret. Each participant has its own (cognitive) similarity scale and distribution, i.e., the similarity value for the most and least similar concept differs between participants. For instance, the absolute values for the concept *River* range from 0.93 to 0.73 for participants that had chosen *River* to be the most similar concept to *Canal*. Overall, SIM-DL values are close to the (weighted) mean similarity judgments, but tend to overestimate.

While these results look promising, the interquartile ranges raise some questions. This becomes especially apparent in case of *Irrigation Canal* and partly also for *Reservoir*. In the first case, while most participants had chosen a high similarity (5^{th} or 6^{th} rank), several subjects ranked *Irrigation Canal* as very dissimilar. There may be two potential explanations for these results. Out of all compared concept descriptions, *Irrigation Canal* is the only one specified as a *non-navigable body of water*, while all others (except *Offshore Platform*) are *navigable*. When subjects compare *Irrigation Canal* to *Canal*, they use the previously made similarity judgments as points of reference. While *Offshore Platform* is too different to serve as a reference, all other concept share a characteristic that is missing for *Irrigation Canal*. In this case, *navigable* becomes the characteristic feature of the set of compared concepts and gets a high weighting. This explanation corresponds well to the variability context weighting proposed by Rodríguez and Egenhofer [143] as well as to Tversky's notion of diagnosticity [168]. Tversky argues that features which are diagnostic for a particular classification have a disproportionate influence on similarity judgments. To study the influence of context on semantic similarity measurement, impact measures such as proposed by Keßler and colleagues can be used [82, 92]. This would allow to determine which kinds of contexts [82] are relevant for a particular test setting. As a result of the test, an extended context model for SIM-DL was developed which is discussed in section 8.4.

A second explanation can be based on different kinds of information processing and extraction. One has to keep in mind that while the similarity server and the participants share the same information about the presented concepts, their representation is different (plaintext versus description logics). The similarity ranking task involves some deductive reasoning steps. For instance, canals were defined as entities which are connected to at least two bodies of water, while rivers have at least one hydrographic feature as origin and one body of water as destination. The underlying ontology represents this using the three relations *connectedTo* and its sub-relations *hasOrigin* and *hasDestination*. When searching for entities connected to waterbodies, an entity with a waterbody as origin satisfies this requirement and should be similar. Participants seem to perform this kind of reasoning, and therefore assign a high rank to *River*. In contrast, irrigation canals have at least one waterbody as origin and one agricultural area as destination. Instead of judging the origin and destination separately, participants may summarize both to a non-matching feature [168].

It is difficult to verify that the reasoning steps specified in the test setting were done by all participants in the same way. The alignment between *constructedAs* and *natural* was probably more evident and, therefore, performed in most cases. One reason could be that there were several natural and constructed feature types and hence the reasoning step was activated more than once. Finally, as the test was not supervised, one cannot rule out the possibility that participants changed their similarity estimation after filling in the feature type names.

The DL concepts (and therefore also the plain text descriptions presented to the participants) specified for the SIM-DL test were defined using roles and their fillers as descriptors. Consequently, the high correlation between human and SIM-DL rankings also support the selection of the multiplicative approach in the roles and fillers test (in section 7.1).

8 CONCLUSIONS AND FUTURE WORK

This chapter summarizes the thesis and points out the most relevant results. Possible extensions to SIM-DL and new application areas are discussed.

8.1 SUMMARY AND ACHIEVED RESULTS

Starting from a list of problems which hinder similarity to support semantics-based (geographic) information retrieval, the thesis describes the development and implementation of the SIM-DL theory and server. The use case of a distributed gazetteer infrastructure (and its Web interface) has been introduced to motivate the need for similarity. First steps towards a geographic feature type ontology have been discussed and specified using the $\mathcal{ALCHQ}$ description logic. Based on a review of previous work and the experience in designing SIM-DL, a framework has been introduced to describe how similarity measures (for information retrieval) work. We argue that this allows to define the semantics of similarity, as the framework clarifies what a particular theory really measures and how the search and target concepts are selected. The similarity ranking obtained using the SIM-DL theory (which is based on the introduced framework) has been compared to human similarity rankings. The theory was implemented as a DIG-compliant similarity server, and the Protégé plug-in demonstrates the ability to integrate the server into the existing semantic Web infrastructure.

The results of the human participants test show a strong and significant correlation between human similarity rankings and the SIM-DL ranking, as well as a significant concordance between the human rankings. Based on these results we cannot falsify the hypothesis, and hence accept it. Summing up, the developed theory and implementation fulfill all the requirements defined in the problem statement.

Has the overall vision of establishing similarity reasoning as a tool for the (geospatial) semantic Web been reached? — Partly. The implementation and integration of SIM-DL is only a first step into this direction. There are several central aspects which need further improvement. For instance, up to now, the computation of similar concepts does not scale and requires more advanced optimization techniques such as caching and approximation (see also [171] for the scalability of description logics such as OWL-DL). As pointed out by Möller [125], the expressivity of $\mathcal{ALCHQ}$ may be too limited to develop complex ontologies for the geo-domain. The human participants test also pointed out a strong need for an extended context model. Additionally, the interpretation of similarity rankings (and values) needs further investigation. Finally, more applications are necessary to study how similarity can improve information retrieval and human computer interaction.

8.2 FUTURE EXTENSIONS TO THE SIMILARITY FRAMEWORK

The framework described in this thesis consists of five steps starting with the selection of search and target concepts and ending with the computation of the overall similarity. While the framework improves the understanding and classification of (semantic) similarity measures for information retrieval, it turns out that two steps are missing.

First, the background (e.g., motivation, cognitive capabilities, environment) of the user running the similarity query is of fundamental importance for adjusting the similarity functions. Up to now, the framework specifies where these adjustments should be defined (on the level of the alignment matrices and the similarity functions[1]; see section 2.2 and figure 20), but not how to derive them from the user's input[2]. Besides the question whether a symmetric or asymmetric measure should be used, or whether the maximum or averaged similarity mode should be applied in case of disjunction, the user's background has an impact on the specification of the compared concepts. As requirement, the framework should therefore specify the application area and intended audience in the first step. Further work is required to figure out how and which information needs to be made explicit.

Second, the framework ends with the determination of the overall similarity without discussing the interpretation of the resulting values (e.g., within a similarity ranking). Even if a theory, such as SIM-DL, focuses on rankings instead of single similarity values, the representation and interpretation of these rankings is not trivial. In fact, the interpretation has to be a mapping from similarity values to representations. The representation, i.e., how the results are displayed to the user, depends on the first step of the extended framework, namely the user's background.

An extended framework consists of the following seven steps:

1. Definition of application area and intended audience
2. Selection of search (query) and target concepts
3. Transformation of concepts to canonical form
4. Definition of an alignment matrix for concept descriptors
5. Application of constructor specific similarity functions
6. Determination of normalized overall similarity
7. Interpretation of the resulting similarity value(s)

One way to deal with these new steps is an extended context model. A first glance at this model is described in section 8.4 and was used for the new versions of the server and Protégé plug-in presented in chapter 6.

---

1 One may also think of cases were asymmetry has impact on the selection of target concepts, e.g., using the *lcs/msc* approach (see section 5.7).

2 This was one of the outcomes from discussions during the *Semantic Similarity and Geospatial Applications* workshop at COSIT 2007.

Finally, the framework can be further extended to support comparable human participants tests. Based on the steps defined by the framework and their implementation within particular similarity theories, it should be possible to derive a set of minimum characteristics which each test design has to implement.

8.3 FUTURE EXTENSIONS TO SIM-DL

This section discusses possible extensions to the SIM-DL similarity theory and server. Most of the proposed extensions will be part of future SIM-DL releases (or have already been integrated into recent versions). Apart from these extensions, the similarity server is still a prototype and future work will focus on performance issues and debugging.

8.3.1 *Circularity, Blocking, and Approximation*

To handle circular definitions[3] such as $C \equiv ... \sqcap (\forall R.C)$ the matrix (and the similarity functions) need to implement a blocking mechanism as known from tableaux algorithms for satisfiability reasoning in DL. For instance, consider the tuple $sim(C, D)$ from the matrix M_1 used to compare a search and target concept (where C is defined as above and $D \equiv ... \sqcap (\forall R.D)$). In order to calculate the similarity between C and D, an alignment matrix M_2 that contains tuples for all possible combinations of the Cartesian product $C \times D$ is created. Since the definition of concept C (and D) is circular, all tuples from M_2 containing $(\forall R.C)$ (and $(\forall R.D)$) will end up in a loop (creating infinite alignment matrices). Instead, such tuples are set as *blocked*. All similarity values for tuples in the matrix M_2 are calculated leaving the blocked tuples aside. The result is an approximated similarity between C and D. Using this value, the blocked tuples can now be computed and M_2 (and finally M_1) can be re-calculated without loops. This tuple-wise blocking often appears in case of negation. If only one part of the tuple is blocked (e.g., if $(\forall R.D)$ is replaced by $(\forall R.E)$), the process continues unfolding E and building matrices until no $\forall$ expression to be compared to $(\forall R.C)$ is left, or its filler is either $\top$ or primitive. As similarity can be computed for this tuple, the value is now used one level (matrix) higher and so on until $sim(C, D)$ can be determined. This kind of blocking is called expression-wise blocking here.

The role of approximation is not bound to overcoming circularity. The measurement of inter-concept similarity turns out to be expensive in terms of computation time. If a threshold is specified for a particular query (which should be the case), future work should focus on how to determine whether a concept will exceed this threshold in the first place, i.e., without running the whole similarity process (see also [13]). For a large set of target concepts, one would first approximate which target concepts are potential matches (exceed the threshold) and compute the exact similarity values for these concepts afterwards. For instance, when measuring similarity (in the maximum similarity mode; see section 5.5.4 and 6.2) between concepts specified using disjunction,

3 The problem of circularity also affects other similarity measures, but was not taken into account so far.

one could return a list of potential candidates (but not a final ranking) without computing the similarity for all pairs of subconcepts. The partial similarity value for potentially matching concepts would be stored and the final similarity (and ranking) could be computed in the following step. The same approach can be used to find potential matches only comparing primitives etc.

Nevertheless, such methods still require to compute the similarity of all subconcepts (and superconcepts, respectively) until the threshold is reached, i.e., if the overall similarity is below the threshold all pairs of superconcepts have to be examined. Turning this process around would make computing similarity less expensive. The similarity value has to be approximated up to a degree which makes it possible to decide whether the target concept can exceed the threshold. If not, the computation can stop at an early stage of the computation process [13]. A candidate for this optimization is the similarity computed for quantifications and restrictions. In fact, it is only necessary to compute the inter-role similarity (which is not expensive in case of $\mathcal{ALCHQ}$; see section 5.5.2) to determine whether a role-filler pair can contribute in a positive way to the overall similarity. The expensive computation of the filler similarity is only necessary if the approximated (leaving the fillers aside) overall similarity exceeds the threshold[4].

As pointed out by Rodríguez [142], one could also try to reason about inter-concept similarity to derive similarities between concepts out of the values computed for other concepts, e.g., their superconcepts. However, similarity is not a transitive relation and highly context dependent which puts serious restrictions on such approaches.

8.3.2 *Taking Instances into Account*

The SIM-DL theory is solely based on concept descriptions to measure similarity. As discussed in section 5.7, information from the ABox may also be relevant for inter-concept similarity. Whether and to which degree depends on the application. In case of the presented gazetteer use case, the features are instances of their respective types, but do not directly correspond to assertions in the ABox (besides pre-given instance-of assertions). The millions of geographic features stored by the ADL gazetteer do not contain the necessary information required for DL instance checking [128]. For example, the Rhein is of type *River*, which is a named concept in our feature type ontology, but the hydrographic features to which the Rhein is connected are not given (as required by the ontological specification of *River*; see section 7.2.2). This gap between ontologies and existing data is a common challenge for semantics-based information retrieval in GIScience. Methods to (semi-automatically) overcome this gap were discussed by Klien [93], Third et al. [164], and Mallenby and Bennett [112]. In addition, if the similarity between concepts is reduced to the ratio of common to distinguishing instances, this may be untransparent and misleading. Imagine that the concepts *Canal* and *River* are compared with respect to a German and a US gazetteer. Even if both gazetteers have the same concept definitions, the similarity between both concepts will be different.

4 For a given threshold t, if $sim_r(R, S) < t$, then $(R.C, S.D) < t$; see appendix A.

To focus on concept descriptions only has the disadvantage that these specifications reflect potential characteristics of instances, not the real distribution of these characteristics (e.g., in case of number restrictions, such as $(\geq 2 connectedTo.HydrographicFeature)$). Further work should focus on a combination of ABox and TBox for similarity reasoning.

In analogy to tableaux algorithms used to compute subsumption based on ABox satisfiability, one could also try to compute inter-concept similarity by generating ABoxes and assertions based on the concept definitions. This would allow to reduce inter-concept similarity to inter-instance similarity.

8.3.3 *Salient Feature Selection*

We have introduced the determination of the user's background as important first step of the similarity framework. Consequently, the SIM-DL theory has to implement this step. This makes the question why a user has chosen a specific (search) concept a key to directed [52] and hence meaningful results. If users are querying a gazetteer for geographic feature types similar to *River*, their motivation might be finding waterways or recreation areas. In the first case, the salient features include *Transportation*, in the second case *Swimming*. Consequently, while *Canal* is semantically close in the first case, *Lake* might by a better candidate for the recreation task (see also [143]). Up to now, most similarity measures try to solve such issues by allowing for manual or semi-automatic weighting. However, there is more to salient feature selection than weighting, which harbors the danger of manipulating the results until they fit.

One promising approach would be to extract salient features based on the *lcs* (and *msc*) approach presented by Möller et al. [126], and discussed in section 5.7. The user specifies some geographic features[5] (not types) as reference, and the similarity server returns similar geographic features based on the least common subsumer computed out of the most specific concepts of these reference features. Consequently, all returned (similar) geographic features contain the characteristics which are common to all reference features. An extension to the SIM-DL server based on this approach is currently implemented within the SimCat project.

Another approach, developed by Janowicz and Raubal [84], proposes an affordance-based measure to reduce the number of compared characteristics to those necessary to fulfill a particular task. Humans tend to classify entities with respect to the functionalities they offer for solving specific tasks within a particular environment. Gibson's theory of affordances [50] accounts for this kind of agent-environment interaction, while most similarity measures isolate similarity estimations from their context (e.g., the cognitive capabilities of the agent, or the task to be solved). Instead, these measures focus on structural and static descriptions of the compared entities and types. An affordance-based specification of the context in which similarity is measured would

5 Note that the salient features are features in the terminology of similarity, i.e., concept descriptors or entity characteristics (attributes), while geographic features are real word entities (or their computational representations).

make the results situation-aware, and hence improves their accuracy. The underlying assumption is that types are the more similar the more common functionalities their instances (entities) afford an agent in solving a particular task. This approach is based on a similarity measure for affordances [84], and therefore requires their (ontological) representation. As discussed by Kuhn [99], further work is required to understand how affordances can be described within ontologies.

In addition, an affordance-based measure also leads to a better understanding of how unfamiliar entities are grouped together to so-called ad-hoc categories [10], which has not been explained in terms of similarity so far.

8.3.4 *Analogy and Alignment*

Besides research on context, further investigations on the role of analogy and alignment may be a promising direction of future research. While analogy focuses on a theory-based view on categorization (see also [115]), an extended alignment theory would improve the correlation between SIM-DL and human similarity rankings. Imagine the following TBox fragment:

$$A \equiv (\geq 2\, connectedTo.Waterbody)$$
$$B \equiv (\geq 1\, hasDestination.Waterbody) \sqcap (\geq 1\, hasOrigin.Spring)$$
$$Waterbody \sqsubseteq HydrographicFeature$$
$$Spring \sqsubseteq HydrographicFeature$$
$$hasDestination \sqsubseteq connectedTo$$
$$hasOrigin \sqsubseteq connectedTo$$

If A is the search concept and B the target concept (and we assume asymmetry), SIM-DL would align $(\geq 2\, connectedTo.Waterbody)$ and $(\geq 1\, hasDestination.Waterbody)$ to an alignable difference as introduced in section 2.1.2 and 5.4. This is not necessarily the most plausible solution. One could also infer $(\geq 2\, connectedTo.HydrographicFeature)$ from the definitions of A and B. This would be aligned as alignable commonality, which raises the question how to deal with the remaining information, namely the difference between $Spring$ and $Waterbody$. In addition, the human participants test pointed out that not all participants perform the same kind of inference during concept comparison. Further work should incorporate recent results form cognitive science with respect to inference and refine SIM-DL's alignment matrices.

8.3.5 *Semantic Similarity between Ontologies and Web Services*

In theory, SIM-DL can also be used to measure semantic similarity between ontologies as long as they share the same primitives or syntactical matching (or related approaches such as text mining) is used in addition. Inter-ontology similarity could be defined as a function over the inter-concept similarity of all concepts in the compared ontologies. However, such an approach would not be very effective and expensive in terms of computation time (and memory). Other approaches focus on translating ontologies to graphs and are based on homomorphisms between these graphs [102, 124].

As discussed by Janowicz [81], a modified version of SIM-DL could also be used to compare Web services for similarity as long as they are annotated using OWL-S[6] or WSML(-Core)[7]. Related approaches were discussed by Hau et al. [67] and Klusch et al. [95], respectively. If the overall vision is semi-automatic, ad-hoc chaining of Web services, aspects such as uncertainty [145], data quality [43], and provenance [31] become essential parts of similarity-based information retrieval.

8.4 kinds of contexts and their impact on semantic similarity measurement

This section presents an extended context model as reaction to the human participants test described in chapter 7. The resulting classification is adoptable to most similarity theories, however, we focus on those developed for description logics (e.g., SIM-DL) here. The section is a shortened and modified version of the paper with the same title by Janowicz [82][8].

The benefit of similarity lies in delivering a ranked list of alternatives for a user's query if no exact match is available, while one major shortcoming is that the results do not necessarily fulfill all user requirements [87]. This is mostly caused by a lack of context information. Existing similarity theories either ignore the influence of context information or reduce the notion of context to restricting the domain of discourse (see section 2.1.1). In this section, we argue that there are several kinds of contexts which have to be addressed during similarity measurement. These contexts have impact on both the measurement process and the later interpretation of similarity values. While some contexts can be inferred [172] or explicitly stated by the user, other kinds are difficult to capture. The context types proposed in this work are relevant for similarity measurement, for further classifications from other application areas see [9, 14, 69].

8.4.1 *Kinds of Contexts*

The following six kinds of contexts can be distinguished and have impact on semantic similarity measurement. For formal definitions and a more detailed view on these contexts, as well as an example on their application, readers are referred to [82]. Future work should focus on how to integrate these contexts into the measurement process and how to quantify their impact on the resulting similarity values [92].

User Context

The first kind of context underlying every information retrieval task is the user context which corresponds to the first step of the extended similarity framework as described in section 8.2. It describes the user's

6 http://www.daml.org/services/owl-s/
7 http://www.wsmo.org/2004/d16/d16.7/v0.1/20040823/
8 Note that the missleading name *Representation Context* used in the original paper was replaced with the term *Specification Context* to avoid confusion with respect to the results of the introduced interpretation function.

cognitive capabilities and cultural background, the current environment, and the user's motivation for using an information retrieval system [69, 146]. There are strong clues from cognitive science that similarity judgments depend on previous knowledge as well as age [97]. For instance, children tend to a perception driven similarity while adults tend towards so-called theory driven similarity [115]. Recent studies from Mark and colleagues point out that similarity also depends on cultural background and language [116]. The influence of the user's (comparison) environment has been examined by Goldstone et al. [57]. Finally, one clearly needs to distinguish between the user's motivation and the query typed into an information retrieval system. If a user is searching for rivers or similar entities, this does neither answer the question of why nor how the data will be used. While the user's capabilities, cultural background, environment, and motivation influence similarity, their impact is difficult to measure (at least from a computer science point of view).

Noise and Intended Context

Based on the definition of context as additional information influencing similarity, one has to distinguish between intended and undesired context. We assume that noise, i.e., undesired context, is the part of the user context that is not formally represented within a context-aware similarity measure. The term noise is chosen here, because this kind of context has impact on human similarity judgments while it is not accessible for computational similarity theories. This results in (apparently random) deviations between human and machine-based similarity judgments. The problem of noise is especially important in case of human participants tests. For instance, when comparing pictures, participants do not only use the depicted entities (*intended stimuli*) for comparison, but also aspects such as the size of the picture or background, e.g., a cloudy sky (see also section 7.1.4). In contrast, intended context is what we are trying to take into account when developing similarity theories and reasoning services (independent of whether we are able to catch all this information).

Application Context

Measuring semantic similarity does not end in itself, but is used to solve a particular task. As argued by Goodman [59] and Medin et al. [123], there is no global law stating how similarity measurement works and what it measures. In implementing specific similarity functions, each application defines the semantics of similarity (values) with respect to its own application area (see section 2.2). We define the application context as additional parameters which the user can pass to the application to influence the way similarity is measured.

For instance, Rodríguez' asymmetric MDSM [143] allows the user to chose between a commonality or variability weighting to elevate the role of specific concept descriptors (see section 2.1.1). In contrast, SIM-DL distinguishes between average and maximum similarity and additionally allows the user to decide whether the measure should be symmetric or not. It is also possible to define a threshold as minimum similarity of interest (see section 6.2). In case of (mobile) recommendation systems, one may also think of K.O. criteria (on instance level) such as a price limit or duration specified by the user (see also [39]).

Besides such explicitly stated information, parts of the application context can be inferred from the user's behavior or spatio-temporal constraints. Daytime and opening hours are classical examples, but user profiles would allow for additional information. One has to keep in mind that the limiting factor is not how much context information can be collected about the user's behavior and motivation[9], but whether it can be incorporated into the similarity measure (e.g., through weights or salient feature selection) and whether it plays a significant role (i.e., has a clear impact on the resulting similarity values [92]).

The application context is the part of the intended context which is captured by the application. A particular similarity service may take spatio-temporal aspects (and their influence; see section 8.4.1) into account, but fail to support other aspects such as legal restrictions. One can argue that computed similarity judgments correlate the better with human judgments the better the application context approximates the intended context.

Discourse Context

In a typical information retrieval scenario, the user only defines the search concept, while the compared-to (target) concepts depend on the domain of discourse, e.g., the examined ontology. The discourse context defines which concepts are compared to the search concept. Along with similarity measures such as MDSM and SIM-DL, we assume that the user is able to restrict the search to a set of concepts by defining a context concept (see section 2.2). This context concept is either part of the ontology or phrased using a (graphical) interface [87] such as the SIM-DL Protégé plug-in. After reclassification, the discourse context is the set of those target concepts which have the context concept as their least common subsumer; see [102] for more details.

The discourse context does not only define which concepts are selected, but also influences similarity (see section 5.2). In case of SIM-DL, all descriptors defining the context concept are not taken into account for the comparison of search and target concept (as they appear in all target concepts)[10].

Specification Context

While the discourse context defines which concepts are compared, the specification context modifies their descriptors in dependence of the application context. This is comparable to the focus change (*dressing*) introduced by Brézillon [21]. Keßler et al. [92] describe the specification context as a set of rules. Each rule maps from an activation condition to a set of concept modifiers and affected (modified) concepts.

9 Which also raises all kinds of privacy issues.
10 Up to now, SIM-DL only allows for primitives (and their intersections) as context concept. The usage of arbitrary concepts would require more complex substitution operations on DL concepts as proposed by Teege [162] and is left for future work.

If the condition for a particular rule from the specification context is true, the rule gets activated. Every affected concept is modified temporarily by either adding or removing the specified concept descriptors. These descriptors are concepts themselves (see [92] and [82] for further details). In dependence of the application, a condition may be a checkbox in a user interface, a FOL axiom, or information extracted from the user's query (e.g., the user's location).

While it is easy to see that modifying the concepts changes their similarity, a quantification of this change turns out to be difficult (see [92]). To measure the impact of the specification context on similarity is interesting future work, as it would allow to infer which parts of the context are of major importance and which could be left aside.

Interpretation Context

Similarity maps compared concepts to a real number, without stating which descriptors of these concepts differ. As argued in section 5.2, a single similarity value measured between two concepts hides most of the relevant information. Consequently, measures such as SIM-DL focus on similarity rankings. The search concept is compared to a set target concepts and the result is an ordered list with descending similarity values. Therefore, we do not argue that certain similarity values are cognitively plausible, but that the computed ranking correlates with human rankings (see chapter 7 for details). We argue that such a ranking puts a single similarity value in context - namely into the context established by the order of similarity values.

This context is called the interpretation context here. It maps the triple search concept, target concept, similarity value from the set of measured similarities[11] to an interpretation value from the domain of interpretations. In its simplest case, the domain of interpretation is formed by the tuple $\{true, false\}$. Depending on the application area and the pairs of compared concepts, each triple is either mapped to $true$ or $false$. In such a case, the question of whether search and target concepts are similar is answered by *yes* or *no*. With respect to the Web gazetteer interface introduced in chapter 4, similarity values are mapped to font sizes (for visualization) using a logarithmic tag cloud algorithm (see figure 10). The Protégé plug-in shown in section 6.3 additionally supports grouping the target concepts into categories.

The interpretation context does not map an isolated similarity value to another domain, but depends on the set of all measured similarities to the target concepts. For instance, the maximum font size is always assigned to the target concept with the highest similarity (to the search concept), independent of a particular value.

8.4.2 *Future Work on Context*

As for other domains, the *context gap* [9] between the user context and its computational representation is also relevant for similarity measurement. While this mismatch cannot be solved, future work should focus on other contexts to improve the correlation between human and ma-

11 between the search concept and those target concepts defined by the discourse context.

chine similarity judgments. A better understanding of the different types of contexts and their influence allows to improve the accuracy of machine-based similarity ratings and make them situation-aware. In addition, as not all context information can be modeled, one can still examine which information is most relevant and which could be left aside (e.g., using context impact measures [92]). Doing so would also help to differentiate between noise and intended context. Further work is necessary to understand how contexts can be compared and how the context impact can be used to predict change in similarity values. The same kind of testing as introduced in chapter 7 is also necessary for the proposed contexts. This is especially important in case of the interpretation context. Different applications (such as mobile decision support systems) may require their own result visualization and interpretation.

In case of mobile applications, certain context information might be available only at a given time or at a given location. This leads to the question on how to update similarity judgments on-line, which relates to AI planning: How to proceed in the absence of this information and how to interpolate or infer it?

Finally, the relationship between context and similarity is reciprocal. While this section describes how to improve the accuracy of machine-based similarity judgments using context, one could also infer context information out of similarity judgments.

8.5 further application areas

Within this thesis, the potential of semantic similarity measurement has been demonstrated by a use case from gazetteer research. Besides various extensions to SIM-DL, future work should focus on applying the results to additional, more advanced use cases. Further human participants tests should be performed using these use cases to underpin the results discussed in chapter 7. In the following, ideas for further applications are described, and a similarity-based identity assumption service for historical places is proposed to demonstrate how more complex scenarios may benefit from similarity [78].

8.5.1 *A Spectrum of Potential Similarity-Based Applications*

The integration of the similarity server within a Spatial Data Infrastructure (SDI) would be an interesting next step to connect OGC web services to reasoning services (such as similarity) on the Semantic Web. For instance, the SIM-DL server could query a Web Catalog Service for similar geographic feature types if a user's query to a Web Feature Server does not deliver the intended results. A conceptual architecture and worklow for integrating similarity-based information retrieval into an SDI has been discussed by Janowicz et al. [87].

Another aspect not covered in detail within this thesis is how to establish a mapping from the gazetteer features to instances within the respective feature type ontology (see also section 8.3.2 and [93]). While this is not of major importance for similarity measurement as such, it is a requirement for the integration of the SIM-DL similarity server within the proposed distributed gazetteer infrastructure.

Similarity could also be applied during ontology engineering. As subsumption reasoning and satisfiability are used to reclassify ontologies and find contradicting concept specifications, similarity could be used to suggest whether the defined concepts match the engineers intention (and hence act as a *fit for purpose* quality indicator). If, for instance, the similarity between *River* and *Canal* is smaller than between *River* and *Sea* this may indicate that the concept descriptions are biased in a certain way (which may be intended).

8.5.2 *Similarity-Based Identity Assumptions for Historical Places*

The domain of cultural heritage is very heterogeneous. The themes or exhibits that museums and related institutions are concerned with range from history of science to various kinds of art, historical documents, and biodiversity. Accordingly, the number and type of preserved exhibits range from millions of collected organisms to a small set of valuable paintings. Creating and maintaining metadata about exhibits and historical facts in general gets increasingly important for scholars and curators in order to structure, manage, and query their own data. As long as metadata is used for internal workflows only (such as the preparation of an exhibition), each institution may develop and maintain their own schema and representation format. To refine and enrich their own knowledge base, or to answer complex scientific questions, interchange with external sources becomes necessary. Cleaning up the local knowledge base is especially important, because one has to keep in mind that historical knowledge may be vague, incomplete, or even misleading. To support these tasks the Committee on Documentation (CIDOC) provides a well established and standardized core ontology (called CIDOC CRM; ISO 21127) [29], intended to annotate heterogeneous cultural heritage information. The goal is to make data available in a machine-readable format (the Resource Description Framework; RDF), supporting reasoning, knowledge integration, mediation, and interchange. The long-term vision is to publish all annotated datasets through Web services, and therefore, create a shared network (GRID) of interlinked historical information which enables automatic metadata harvesting. The CIDOC conceptual reference model can be regarded as the underlying semantic level, which provides meaning within the intended cultural heritage data infrastructure (which can bee seen analogously to a Spatial Data Infrastructure) by delivering a common metadata schema. Instead of trying to reach a community wide agreement on definitions for concrete entity classes (such as types of exhibits), the strength of CIDOC CRM lies in defining an abstract but interrelated vocabulary. This reference model describes the foundation of historical facts, namely established links (relations) between places, actors, objects, and events [29].

For instance, one could describe the Battle of Trafalgar either using exact geographic coordinates (which were not available at that time) and a timestamp, or by establishing links to other historical facts. The Battle of Trafalgar can be described as an event which took place during the Napoleonic Wars (which are also events) and was carried out between the British fleet and the combined French/Spanish fleet as actors. The naval battleground is also the place where Vice-Admiral Horatio Nelson died.

To make use of external data sources, however, a common language is not sufficient. It must be guaranteed that the collected metadata refers to the same real world phenomenon (which could be a historical place, person, event, or object) as the local datasets. Global authorities, such as the Alexandria Digital Library Gazetteer, provide unique identifiers and annotated datasets for some common kinds of real world phenomena. Scholars can refer to these global identifiers in addition to (or instead of) their local identifiers and, therefore, reduce maintenance effort and redundancy. In addition, this enables data interchange with other institutions. If compared datasets refer to the same global identifier, and the scholar decides to trust the global authority as well as the external party (which linked their dataset to the specific identifier), it can be assumed that the same real world phenomenon is meant.

So far, most datasets do not refer to global authorities and scholars must decide as the case arises whether the harvested information is relevant for their own knowledge base. There are several reasons for this [78]:

- Knowledge about historical places is often vague and incomplete.
- Place names are not unique (even within the same geographic area).
- Place names may refer to cities, rivers, valleys, mountains, etc.
- Place names can be misinterpreted (e.g., 'Al Wahat' meaning oasis).
- Names also refer to varying geopolitical units (e.g., nomads) or prominent (artificial) landmarks (e.g., telegraph stations).
- There are out-dated place or even country names (e.g., the Soviet Union).

The most significant reason why global identifiers can only partially solve the problem of identity is that using gazetteers to determine whether two datasets refer to the same real world place presumes that all involved institutions have manually annotated millions of local datasets beforehand. This is not the case until now. Therefore, an identity assumption assistant should support scholars in analyzing the harvested metadata and return promising data, i.e., external datasets which probably refer to the same real world place addressed by the local data. The identity assumption theory used by such an assistant should be non-rigid in a way that it returns a ranked list of estimations instead of trying to automatically conclude safe predictions from vague historical data.

In practice, disambiguation via gazetteers and other global authorities is often difficult, expensive, and error-prone (especially for subordinate geopolitical units, events, actors, etc.). The theory behind the identity assumption service is to use the links established via the CIDOC

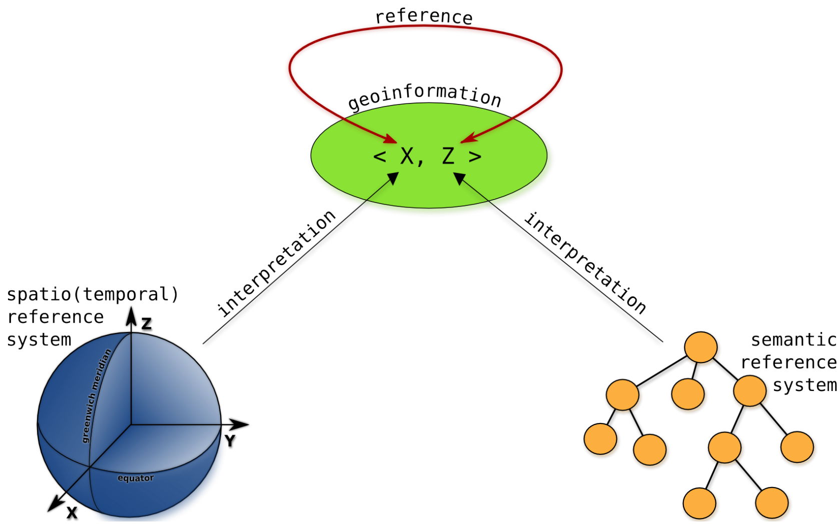

Figure 32: Using thematic information as support for spatio-temporal reference.

CRM annotation between places, actors, objects, and events as additional reference points. In other words, taking Goodchild's geographic reality (geoinformation as a tuple defined by a spatio-temporal location vector and a thematic vector [58]) and Kuhn's notion of semantic reference systems [100, 101] into account, the underlying idea is to use thematic information as support for spatio-temporal reference (see figure 32). The same way as the spatio-temporal location vector is interpreted using a spatio-temporal reference system, thematic information is interpreted by a semantic reference system [100] defined by CIDOC CRM as a formal ontology. Similarity and classical (spatio-temporal and subsumption) reasoning are functions defined for such a reference system.

Using similarity as part of the puzzle of identity assumptions draws the metaphor from a geographical notion of location to the location within a network of historical facts, and the spatial 'next-to' relation to a thematic relation based on similarity assessments [78]. Two datasets probably refer to the same historic place, if both describe the same (or similar) historic events or actors and connect those via the same (or similar) relationships to the described places. Thus, from a similarity point of view, place identity can be expressed as a function of common relation-object tuples (see figure 33). The more common (RDF) triples two instances share, the more similar they are and the higher is the probability that both point to the same real world place. In other words, measuring similarity between real world places means to develop (or apply) distance metrics for their descriptions and to determine their overlap. However, instances within a knowledge base always represent the approximated and partial knowledge an authority or museum has about a real world phenomenon. Hence, even if two instances share all triples, identity cannot be guaranteed.

Each object (such as a particular event, actor, or object) related to the compared place is itself identified by another relation-object tu-

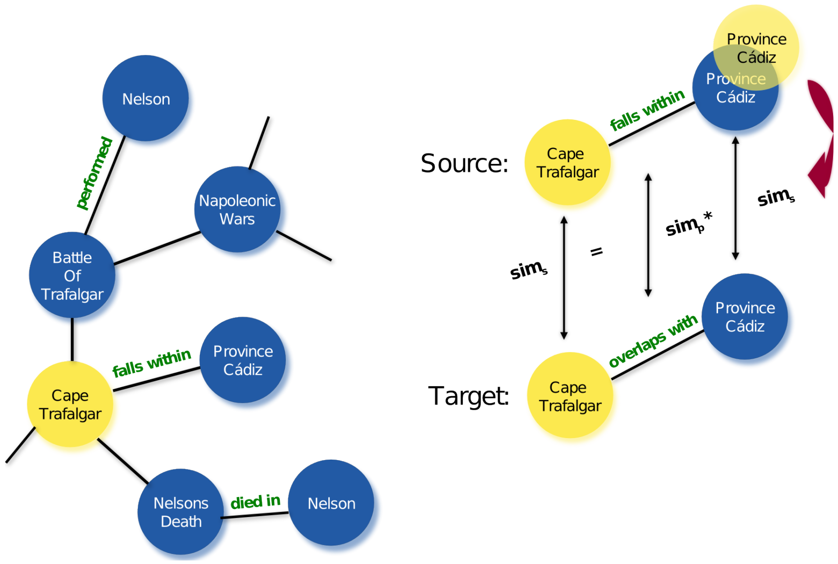

Figure 33: Similarity-based identity assumptions for historical places.

ple. Therefore, to determine place identity, all descriptions have to be compared for similarity until only literals are left for syntactic matchmaking (see [78] for details). Figure 33 illustrates this process for the Battle of Trafalgar scenario introduced above. sim_s is the similarity of compared subjects (and those objects which are subjects of additional RDF tiples), while sim_p is the relation (RDF predicate) similarity.

Parts of the similarity theory presented in this thesis may act as starting point to develop such an identity assumption service in the future.

APPENDIX A

Chapter 5 introduces the similarity functions used for SIM-DL without pointing out how they are chained together to measure overall similarity between complex concepts. This appendix shows the used functions. Further details and explanations of the symbols are given in section 5.3 and 3.2, as well as [79, 85].

Note that the sets of tuples selected by the alignment matrix (see section 5.4) are represented by the letter S followed by an abbreviation for the type of constructor. For instance, SI is the set of concepts on union level of C where each C_i is formed by intersection.

sim_{u_w} is the weighted sum of similarities for all tuples (C_i, D_j). The weighting ω ($\sum \omega_{ij} = 1$) can be either determined by the count of tuples or by analyzing the ontological structure [79]. If the similarity of a particular tuple is 1, $sim_u = 1$. Per default, SIM-DL uses the maximum similarity mode with the sim_{u_w} function (see section 5.5.4).

$$sim_{u_m}(C, D) = max(sim_i(C_i, D_j)); \text{ where } (C_i, D_j) \in SI \tag{A.1}$$

$$sim_{u_w}(C, D) = \sum_{(C_i, D_j) \in SI} \omega_{ij} * sim_i(C_i, D_j) \tag{A.2}$$

Following the $\mathcal{ALCHQ}$ canonical normal form (see section 5.3), each C_i (respectively D_j) is an intersection of primitives or concepts formed by restrictions or quantifications. sim_i is the function that determines similarity on this level as normalized sum derived from the similarity functions for the involved constructors. The normalization factor σ is defined as the sum of cardinalities derived from the sets of compared tuples (SP, SE, SF, $SMIN$ and $SMAX$).

$$\begin{aligned} sim_i(C, D) = \\ \frac{1}{\sigma} \Bigg(& \sum_{(A,B) \in SP} sim_p(A, B) + \sum_{(R,S) \in SE} sim_e(exists_R(C), exists_S(D)) \\ & + \sum_{(R,S) \in SF} sim_f(forall_R(C), forall_S(D)) + \sum_{(R,S) \in SMIN} sim_m(min_R(C), min_S(D)) \\ & + \sum_{(R,S) \in SMAX} sim_m(max_R(C), max_S(D)) \Bigg) \end{aligned} \tag{A.3}$$

Primitives have no description that can be compared, hence Jaccard's coefficient is used to determine their similarity. Primitives are the more similar, the more complex concepts (within the context) are subsumed by both.

$$sim_p(A,B) = \frac{\mid \{C \mid C \sqsubset A) \wedge (C \sqsubset B)\} \mid}{\mid \{C \mid C \sqsubset A) \vee (C \sqsubset B)\} \mid} \quad \text{(A.4)}$$

sim_e compares concepts formed by existential quantifications. The similarity is the product of role and filler similarity. The second sum (see sim_i) is necessary as there may be more than one existential quantification for the same role.

$$sim_e(exists_R(C), exists_S(D)) = sim_r(R,S) * \sum_{(C'_i,D'_j)\in SE} sim_u(C'_i), D'_j)) \quad \text{(A.5)}$$

sim_f compares concepts formed by value restriction. The similarity is the product of role and filler similarity.

$$sim_f(forall_R(C), forall_S(D)) = sim_r(R,S) * sim_u(forall_R(C), forall_S(D)) \quad \text{(A.6)}$$

The similarity (sim_m) between concepts formed by quantified number restrictions is the product of the similarities determined for the involved roles, fillers and their maximal or minimal occurrence (cardinality). sim_m is used as an abbreviation here, in fact minimum and maximum restrictions are handled separately (i.e. m is replaced by min respectively max). The normalization $m_{RS}(total)$ is the highest maximum (respectively minimum) restriction for R or S within the context. If one cardinality is explicitly set to 0 (while the other is not), $sim_m = 0$.

$$sim_m(C,D) = sim_r(R,S) * \left(1 - \frac{\mid m_R(C) - m_S(D) \mid}{m_{RS}(total)}\right) * sim_u(E,F) \quad \text{(A.7)}$$

The similarity between roles (sim_r) is their normalized distance within the hierarchy. The normalization is depth-dependent to indicate that the distance from node to node decreases with increasing depth of R and S within the hierarchy.

$$sim_r(R,S) = \frac{depth(lub(R,S))}{depth(lub(R,S)) + edge\_distance(R,S)} \quad \text{(A.8)}$$

If roles are not organized within a hierarchy but within a neighborhood, sim_n is used for comparison.

$$sim_n(R,S) = \frac{max\_distance_n - edge\_distance(R,S)}{max\_distance_n} \quad \text{(A.9)}$$

BIBLIOGRAPHY

[1] ANSI/NISO. *ANSI/NISO Z39.19 - 2005 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. NISO Press, 2005. (Cited on page 32.)

[2] R. Araújo and H. S. Pinto. Semilarity: Towards a model-driven approach to similarity. In *International Workshop on Description Logics (DL2007)*, volume 20, pages 155–162, Bozen-Bolzano, Italy, June 2007. Bolzano University Press. (Cited on pages 1, 8, and 25.)

[3] R. Araújo and H. S. Pinto. Towards semantics-based ontology similarity. In P. Shvaiko, J. Euzenat, F. Giunchiglia, and B. He, editors, *Proceedings of the Workshop on Ontology Matching (OM2007) at the International Semantic Web Conference (ISWC2007)*, Busan, South Korea, November 2007. (Cited on pages 1, 8, and 25.)

[4] F. G. Ashby and N. A. Perrin. Toward a unified theory of similarity and recognition. *Psychological Review*, 95:124–150, 1988. (Cited on pages 17 and 57.)

[5] F. Baader and U. Sattler. An overview of tableau algorithms for description logics. *Studia Logica*, 69:5–40, 2001. (Cited on page 25.)

[6] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003. ISBN 0-521-78176-0. (Cited on pages 24, 37, 54, and 65.)

[7] F. Baader, R. Küsters, and F. Wolter. *The Description Logics Handbook: Theory, Implementations, and Applications*, chapter Extensions to Description Logics. Cambridge University Press, 2003. (Cited on page 59.)

[8] F. Baader, B. Sertkaya, and A.-Y. Turhan. Computing the least common subsumer w.r.t. a background terminology. *Journal of Applied Logic*, 5:392–420, 2007. (Cited on page 60.)

[9] L. Barkhuus. *The Context Gap: An Essential Challenge to Context-Aware Computing*. Ph.D. Thesis, IT University of Copenhagen, December 2004. (Cited on pages 99 and 102.)

[10] L. Barsalou. Ad hoc categories. *Memory & Cognition*, 11:211–227, 1983. (Cited on pages 8 and 98.)

[11] L. Barsalou. Situated simulation in the human conceptual system. *Language and Cognitive Processes*, 5(6):513–562, 2003. (Cited on page 8.)

[12] L. Barsalou, W. Yeh, B. Luka, K. Olseth, K. Mix, and L. Wu. Concepts and meaning. In K. Beals, G. Cooke, K. Kathman, K. McCullough, S. Kita, and D. Testen, editors, *Papers from the parasession on conceptual representations*, volume 29 of *Chicago Linguistics Society*, pages 23–61. University of Chicago: Chicago Linguistics Society, 1993. (Cited on page 8.)

[13] R. J. Bayardo, Y. Ma, and R. Srikant. Scaling up all pairs similarity search. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pages 131–140, Banff, Alberta, Canada, May 2007. ACM. ISBN 978-1-59593-654-7. (Cited on pages 1, 95, and 96.)

[14] M. Bazire and P. Brézillon. Understanding context before using it. In A. K. Dey, B. N. Kokinov, D. B. Leake, and R. M. Turner, editors, *Modeling and Using Context, 5th International and Interdisciplinary Conference (CONTEXT 2005)*, volume 3554 of *Lecture Notes in Computer Science*, pages 29–40. Springer, July 2005. (Cited on pages 14 and 99.)

[15] S. Bechhofer. The dig description logic interface. In *International Workshop on Description Logics (DL2003)*, Rome, Italy, September 2003. (Cited on pages 3, 25, 26, and 65.)

[16] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P. F. Patel-Schneider, and L.A. Stein. Owl web ontology language reference. W3c recommendation, W3C, February 2004. (Cited on page 24.)

[17] A. Bernstein, E. Kaufmann, C. Kiefer, and C. Bürki. Simpack: A generic java library for similarity measures in ontologies. Technical report, Department of Informatics, University of Zurich, 2005. (Cited on page 12.)

[18] A. Borgida, T. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In *International Workshop on Description Logics (DL2005)*, volume 147 of *CEUR Workshop Proceedings*. CEUR, Edinburgh, Scotland, UK, July 2005. (Cited on pages 1, 8, 25, and 60.)

[19] J. Bortz and R. Weber. *Statistik für Human- und Sozialwissenschaftler*. Springer-Lehrbuch. Springer Medizin, Heidelberg, 2005. (Cited on pages 3, 81, 83, and 89.)

[20] S. Brandt, R. Küsters, and A.-Y. Turhan. Approximation and difference in description logics. In D. Fensel, F. Giunchiglia, D. L. McGuiness, and M.-A. Williams, editors, *Proceedings of the 8th International Conference on Principles of Knowledge Representation and Reasoning (KR2002)*, pages 203–214, San Francisco, USA, April 2002. Morgan Kaufman. (Cited on pages 25, 40, 55, and 59.)

[21] J. Brézillon and P. Brézillon. Context modeling: Context as a dressing of a focus. In B. N. Kokinov, D. C. Richardson, T. Roth-Berghofer, and L. Vieu, editors, *Modeling and Using Context, 6th International and Interdisciplinary Conference (CONTEXT*

2007), volume 4635 of *Lecture Notes in Computer Science*, pages 136–149, Copenhagen, Denmark, August 2007. Springer. (Cited on page 101.)

[22] H.T. Bruns and M.J. Egenhofer. Similarity of spatial scenes. In J. M. Kraak and M. Moleenar, editors, *Proceedings of the 7th International Symposium on Spatial Data Handling (SDH'96)*, pages 173–184, Delft, The Netherlands, August 1996. Taylor and Francis. (Cited on pages 55 and 77.)

[23] P. Burmeister. Formal concept analysis with conimp: Introduction to the basic features. Technical report, Department of Mathematics (Arbeitsgruppe Allgemeine Algebra und Diskrete Mathematik), Darmstadt University of Technology, Darmstadt, Germany, 2003. (Cited on page 62.)

[24] R. Casati and A. C. C. Varzi. *Parts and Places. The Structures of Spatial Representation*. MIT Press [Bradford Books], Cambridge and London, 1999. (Cited on page 28.)

[25] H. Chen, T. Finin, and A. Joshi. An ontology for context-aware pervasive computing environments. *Knowledge Engineering Review*, 18(3):197–207, 2003. (Cited on page 13.)

[26] S.M. Chen, M.S. Yeh, and P.Y. Hsiao. A comparison of similarity measures of fuzzy values. *Fuzzy Sets and Systems*, 72(1):79–89, 1995. (Cited on pages 37 and 61.)

[27] A. Cohn. Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica*, 1(3):275–316, 1997. (Cited on pages 55 and 77.)

[28] A. Cregan, R. Schwitter, and T. Meyer. Sydney owl syntax - towards a controlled natural language syntax for owl 1.1. In *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions, Innsbruck*, volume 258, Innsbruck, Austria, June 2007. CEUR-WS. (Cited on page 86.)

[29] N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff. *Definition of the CIDOC Conceptual Reference Model (Manual Version 4.2.1)*, October 2006. (Cited on page 104.)

[30] V. Cross and T. Sudkamp. *Similarity and Computability in Fuzzy Set Theory: Assessments and Applications*, volume 93 of *Studies in Fuzziness and Soft Computing*. Physica-Verlag, 2002. (Cited on pages 17 and 61.)

[31] P. P. da Silva, D. L. McGuinness, and R. McCool. Knowledge provenance infrastructure. *IEEE Data Engineering Bulletin*, 26(4): 26–32, 2003. (Cited on page 99.)

[32] 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, June 2005. (Cited on pages 1, 8, 17, 25, 37, 58, and 59.)

[33] C. d'Amato, N. Fanizzi, and F. Esposito. A dissimilarity measure for $\mathcal{ALC}$ concept descriptions. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC2006)*, pages 1695–1699, Dijon, France, April 2006. ACM. (Cited on pages 1, 8, 17, 25, 37, 58, and 59.)

[34] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation? *AI Magazine*, 14(1):17–33, 1993. (Cited on page 22.)

[35] I. Dickinson. Implementation experience with the dig 1.1 specification, hpl-2004-85. Technical report, Hewlett-Packard, 2004. (Cited on page 65.)

[36] D. Dubois and H. Prade. A unifying view of comparison indices in a fuzzy set-theoretic framework. In R.R. Yager, editor, *Recent Development in Fuzzy Set and Possibility Theory*, pages 3–13. Pergamon Press, 1982. (Cited on pages 56 and 61.)

[37] M. Egenhofer and K. Al-Taha. Reasoning about gradual changes of topological relationships. In A. Frank, I. Campari, and U. Formentini, editors, *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, volume 639 of *Lecture Notes in Computer Science*, pages 196–219. Springer, Heidelberg-Berlin, 1992. (Cited on pages 55 and 77.)

[38] M. Egenhofer and R. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5 (2):161–174, 1991. (Cited on pages 55 and 77.)

[39] M. Espeter. Location-based decision support for user groups (diploma thesis), 2007. (Cited on pages 2 and 101.)

[40] B. Falkenhainer, K.D. Forbus, and D. Gentner. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41:1–63, 1989. (Cited on pages 11, 15, 73, and 83.)

[41] K. Forbus, D. Gentner, A. B. Markman, and R. W. Ferguson. Analogy just looks like high level perception: Why a domain-general approach to analogical mapping is right. *Journal of Experimental & Theoretical Artificial Intelligence*, 10(2):231–257, 1998. (Cited on page 83.)

[42] A. Formica. Ontology-based concept similarity in formal concept analysis. *Information Sciences*, 176(18):2624–2641, September 2006. (Cited on page 62.)

[43] A. Frank. Data quality ontology: An ontology for imperfect knowledge. In S. Winter, M. Duckham, L. Kulik, and B. Kuipers, editors, *Spatial Information Theory, 8th International Conference (COSIT 2007)*, volume 4736 of *Lecture Notes in Computer Science*, pages 406–420, Melbourne, Australia, September 2007. Springer. (Cited on page 99.)

[44] A. U. Frank. Similarity measures for semantics: What is observed? In *COSIT'07 Workshop on Semantic Similarity Measurement and Geospatial Applications*, Melbourne, Australia, September 2007. (Cited on page 7.)

[45] C. Freksa. Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54(1):199–227, 1992. (Cited on page 55.)

[46] M. Gahegan, R. Agrawal, T. Banchuen, and D. DiBiase. Building rich, semantic descriptions of learning activities to facilitate reuse in digital libraries. *International Journal on Digital Libraries*, 7(1):81–97, 2007. (Cited on page 8.)

[47] B. Ganter, G. Stumme, and R. Wille, editors. *Formal Concept Analysis, Foundations and Applications*, volume 3626 of *Lecture Notes in Computer Science*, 2005. Springer. ISBN 3-540-27891-5. (Cited on page 62.)

[48] P. Gärdenfors. *Conceptual Spaces - The Geometry of Thought*. Bradford Books, MIT Press, Cambridge, MA, 2000. (Cited on page 8.)

[49] D. Gentner and K. D. Forbus. MAC/FAC: A model of similarity-based retrieval. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, pages 504–509. Erlbaum, 1991. (Cited on pages 11, 15, 73, and 83.)

[50] J. Gibson. The theory of affordances. In R. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing - Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Ass., Hillsdale, New Jersey, 1977. (Cited on pages 9, 73, and 97.)

[51] R. L. Goldstone. Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:3–28, 1994. (Cited on pages 1, 10, and 11.)

[52] R. L. Goldstone. Mainstream and avant-garde similarity. *Psychologica Belgica*, 35:145–165, 1995. (Cited on page 97.)

[53] R. L. Goldstone. Hanging together: A connectionist model of similarity. In Grainger J. and A. M. Jacobs, editors, *Localist Connectionist Approaches to Human Cognition*, pages 283 – 325. Lawrence Erlbaum Associates, 1998. (Cited on pages 10, 11, and 15.)

[54] R. L. Goldstone and A. Kersten. Concepts and categorization. In A. Healy and R. Proctor, editors, *Comprehensive Handbook of Psychology*, volume 4, pages 599–621. 2003. (Cited on page 58.)

[55] R. L. Goldstone and D. Medin. Similarity, interactive activation, and mapping: An overview. In K. Holyoak and J. Barnden, editors, *Analogical Connections: Advances in Connectionist and Neural Computation Theory*, volume 2, pages 321–362. Ablex, Norwood, USA, 1994. (Cited on page 10.)

[56] R. L. Goldstone and J. Son. Similarity. In K. Holyoak and R. Morrison, editors, *Cambridge Handbook of Thinking and Reasoning*, pages 13–36. Cambridge University Press, Cambridge, 2005. (Cited on pages xiii, 1, 7, 10, 11, 17, 42, 56, and 57.)

[57] R. L. Goldstone, D. L. Medin, and J. Halberstadt. Similarity in context. *Memory & Cognition*, 25:237–255, 1997. (Cited on pages 10, 43, and 100.)

[58] M. Goodchild. Geographical data modeling. *Computers and Geosciences*, pages 401–408, 1992. (Cited on page 106.)

[59] N. Goodman. Seven strictures on similarity. In *Problems and projects*, pages 437–447. Bobbs-Merrill, New York, 1972. (Cited on pages 7, 13, 16, and 100.)

[60] R. A. M. Gregson. *Psychometrics of similarity*. Academic Press, New York, 1975. (Cited on page 16.)

[61] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993. (Cited on page 19.)

[62] N. Guarino. Understanding, building, and using ontologies: A commentary to "using explicit ontologies in kbs development", by van heijst, schreiber, and wielinga. *International Journal of Human and Computer Studies*, 46:293–310, 1997. (Cited on pages 21 and 22.)

[63] N. Guarino. Formal ontology and information systems. In N. Guarino, editor, *International Conference on Formal Ontology in Information Systems (FOIS1998)*, pages 3–15. IOS Press, Trento, Italy, 1998. (Cited on pages 19, 21, 22, 23, and 30.)

[64] N. Guarino. Some ontological principles for designing upper level lexical resources. In Rubio A., N. Gallardo, R. Castro, and A. Tejada, editors, *Proceedings of First International Conference on Language Resources and Evaluation. ELRA - European Language Resources Association*, pages 527–534, Granada, Spain, May 1998. (Cited on page 19.)

[65] U. Hahn, N. Chater, and L. B. Richardson. Similarity as transformation. *Cognition*, 87:1–32, 2003. (Cited on page 1.)

[66] S. M. Harrison. A comparison of still, animated, or nonillustrated on-line help with written or spoken instructions in a graphical user interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 82–89, Denver, Colorado, USA, May 1995. ACM Press/Addison-Wesley Publishing Co. New York, NY, USA. (Cited on page 85.)

[67] J. Hau, W. Lee, and J. Darlington. A semantic similarity measure for semantic web services. In *Web Service Semantics: Towards Dynamic Business Integration Workshop at the World Wide Web Conference 2005*, Chiba, Japan, May 2005. (Cited on page 99.)

[68] B. Hecht and M. Martin Raubal. Geosr: Geographically explore semantic relations in world knowledge. In L. Bernard, A. Friis-Christensen, and H. Pundt, editors, *11th AGILE International Conference on Geographic Information Science (AGILE 2008)*, Lecture Notes in Geoinformation and Cartography, pages 95–113, Girona, Spain, May 2008. Springer. (Cited on page 83.)

[69] K Henricksen and J. Indulska. Modelling and using imperfect context information. In *Proceedings of the Second IEEE International Conference on Pervasive Computing and Communications. Workshop on Context Modelling and Reasoning (CoMoRea'04)*, pages 33–37, Orlando, Florida, USA, March 2004. IEEE Computer Society. (Cited on pages 14, 99, and 100.)

[70] L. L. Hill. *Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing)*. The MIT Press, 2006. ISBN 026208354X. (Cited on page 27.)

[71] L. L. Hill, J. Frew, and Z. Qi. Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, Vol. 5, No. 1, 1999. (Cited on page 33.)

[72] I. Horrocks. *The description logic handbook: theory, implementation, and applications*, chapter Implementation and optimization techniques, pages 306–346. Cambridge University Press, New York, NY, USA, 2003. ISBN 0-521-78176-0. (Cited on pages 39, 40, 65, and 87.)

[73] I. Horrocks. Applications of description logics: State of the art and research challenges. In F. Dau, M.-L. Mugnier, and G. Stumme, editors, *Conceptual Structures: Common Semantics for Sharing Knowledge, 13th International Conference on Conceptual Structures (ICCS 2005)*, volume 3596 of *Lecture Notes in Computer Science*, pages 78–90, Kassel, Germany, July 2005. Springer. (Cited on page 24.)

[74] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen. From shiq and rdf to owl: The making of a web ontology language. *Journal of Web Semantics*, 1(1):7–26, 2003. ISSN 1570-8268. (Cited on page 65.)

[75] IHO. *International Hydrographic Organization Transfer Standard for Digital Hydrographic Data S-57 Edition 3.1*. International Hydrographic Bureau, 2000. (Cited on page 85.)

[76] G. Janée. Rethinking gazetteers and interoperability. In *International Workshop on Digital Gazetteer Research & Practice*, Santa Barbara, California, USA, December 2006. (Cited on page 29.)

[77] K. Janowicz. Extending semantic similarity measurement with thematic roles. In A. M. Rodríguez, I. F. Cruz, M. Egenhofer, and S. Levashkin, editors, *First International Conference on GeoSpatial Semantics (GeoS 2005)*, volume 3799 of *Lecture Notes in Computer Science*, pages 137–152, Mexico City, Mexico, November 2005. Springer. (Cited on pages 4, 10, and 60.)

[78] K. Janowicz. Towards a similarity-based identity assumption service for historical places. In M. Raubal, H. J. Miller, A. U. Frank, and M. F. Goodchild, editors, *Geographic Information Science, 4th International Conference (GIScience 2006)*, volume 4197 of *Lecture Notes in Computer Science*, pages 199–216, Münster, Germany, September 2006. (Cited on pages 1, 4, 12, 27, 103, 105, 106, and 107.)

[79] K. Janowicz. Sim-DL: Towards a semantic similarity measurement theory for the description logic $\mathcal{ALCNR}$ in geographic information retrieval. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, OTM Confederated International Workshops and Posters, AWeSOMe, CAMS, COMINF, IS, KSinBIT, MIOS-CIAO, MONET, OnToContent, ORM, PerSys, OTM Academy Doctoral Consortium, RDDS, SWWS, and SeBGIS. Proceedings, Part II*, volume 4278 of *Lecture Notes in Computer Science*, pages 1681 – 1692. Springer, Montpellier, France, October 2006. (Cited on pages 1, 4, 8, 13, 24, 25, 46, 50, and 109.)

[80] K. Janowicz. Similarity-based retrieval for geospatial semantic web services specified using the web service modeling language (wsml-core). In *Terra Cognita - Directions to the Geospatial Semantic Web Workshop, Workshop at the International Semantic Web Conference 2006*, Athens, USA, November 2006. (Cited on pages 4, 46, and 75.)

[81] K. Janowicz. Similarity-based retrieval for geospatial semantic web services specified using the web service modeling language (wsml-core). In A. Scharl and K. Tochtermann, editors, *The Geospatial Web - How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society*, Lecture Notes in Computer Science, pages 235–245. Springer, Berlin, 2007. (Cited on pages 4, 5, 13, 73, 75, and 99.)

[82] K. Janowicz. Kinds of contexts and their impact on semantic similarity measurement. In *5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea'08) at the 6th IEEE International Conference on Pervasive Computing and Communication (PerCom'08)*, Hong Kong, March 2008. IEEE Computer Society. (Cited on pages 5, 8, 90, 99, and 102.)

[83] K. Janowicz and C. Keßler. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 10(22):1129–1157, 2008. (Cited on pages 4, 19, 28, 30, 31, 32, 34, and 35.)

[84] K. Janowicz and M. Raubal. Affordance-based similarity measurement for entity types. In S. Winter, M. Duckham, L. Kulik, and B. Kuipers, editors, *Spatial Information Theory, 8th International Conference (COSIT 2007)*, volume 4736 of *Lecture Notes in Computer Science*, pages 133–151, Melbourne, Australia, September 2007. Springer. (Cited on pages 5, 8, 58, 97, and 98.)

[85] 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 F. T. Fonseca, A. Rodríguez, and S. Levashkin, editors, *Second International Conference on GeoSpatial Semantics (GeoS 2007)*, number 4853 in Lecture Notes in Computer Science, pages 128–145, Mexico City, Mexico, November 2007. Springer. (Cited on pages 1, 4, 5, 8, 13, 25, 29, 30, 35, and 109.)

[86] 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 L. Bernard, A. Friis-Christensen, and H. Pundt, editors, *11th AGILE International Conference on Geographic Information Science (AGILE 2008)*, Lecture Notes in Geoinformation and Cartography, pages 115–133, Girona, Spain, 5-8. May 2008. Springer. (Cited on pages 5, 73, 74, 84, and 89.)

[87] K. Janowicz, M. Wilkes, and M. Lutz. Similarity-based information retrieval and its role within spatial data infrastructures. In *5th International Conference on Geographic Information Science (GIScience 2008)*, Park City, Utah, USA, September 2008; forthcoming. (Cited on pages 30, 34, 99, 101, and 103.)

[88] M. Jones and B. C. Love. Beyond common features: The role of roles in determining similarity. *Cognitive Psychology*, 55:196–231, 2007. (Cited on page 11.)

[89] I. Jurisica. Dkbs-tr-94-5: Context-based similarity applied to retrieval of relevant cases. Technical report, University of Toronto, Department of Computer Science, Toronto, 1994. (Cited on pages 13 and 38.)

[90] M. G. Kendall and B. Babington Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939. (Cited on page 83.)

[91] C. Keßler. Similarity measurement in context. In B. M. Kokinov, D. C. Richardson, T. Roth-Berghofer, and L. Vieu, editors, *6th International and Interdisciplinary Conference, CONTEXT 2007*, volume 4635, pages 277–290. Springer, Copenhagen, Denmark, August 2007. (Cited on pages 4, 7, 13, and 14.)

[92] C. Keßler, M. Raubal, and K. Janowicz. The effect of context on semantic similarity measurement. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops, OTM Confederated International Workshops and Posters, AWeSOMe, CAMS, OTM Academy Doctoral Consortium, MONET, OnToContent, ORM, PerSys, PPN, RDDS, SSWS, and SWWS 2007, Proceedings, Part II*, number 4806 in Lecture Notes in Computer Science, pages 1274–1284, Vilamoura, Portuga, November 2007. Springer. (Cited on pages 4, 5, 13, 39, 71, 90, 99, 101, 102, and 103.)

[93] E. Klien. A rule-based strategy for the semantic annotation of geodata. *Transactions in GIS, Special Issue on the Geospatial Semantic Web*, 11(3):437–452, 2007. (Cited on pages 96 and 103.)

[94] A.. Klippel, M. Worboys, and M. Duckham. Identifying factors of geographic event conceptualization. *International Journal of Geographical Information Science*, (accepted for publication). (Cited on page 8.)

[95] M. Klusch, B. Fries, and K. P. Sycara. Automated semantic web service discovery with owls-mx. In H. Nakashima, M. P. Wellman, G. Weiss, and P. Stone, editors, *5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, pages 915–922, Hakodate, Japan, May 2006. ACM. (Cited on page 99.)

[96] H. Knublauch, R.W. Fergerson, N.F. Noy, and M.A. Musen. The Protégé OWL plugin: An open development environment for semantic web applications. In *Third International Semantic Web Conference (ISWC2004)*, number 3298 in Lecture Notes in Computer Science, pages 229–243, Hiroshima, Japan, November 2004. Springer. (Cited on page 69.)

[97] L. Kotovsky and D. Gentner. Comparison and categorization in the development of relational similarity. *Child Development*, 67: 2797–2822, 1996. (Cited on page 100.)

[98] C. L. Krumhansl. Concerning the applicability of geometric models to similarity data: the interrelationship between similarity and spatial density. *Psychological Review*, 85:445–463, 1978. (Cited on page 57.)

[99] W. Kuhn. An image-schematic account of spatial categories. In S. Winter, M. Duckham, L. Kulik, and B. Kuipers, editors, *Spatial Information Theory, 8th International Conference (COSIT 2007) Proceedings*, volume 4736 of *Lecture Notes in Computer Science*, pages 152–168, Melbourne, Australia, September 2007. Springer. (Cited on page 98.)

[100] W. Kuhn. Semantic reference systems (guest editorial). *International Journal of Geographical Information Science*, 17(5):405–409, 2003. (Cited on pages 21, 22, and 106.)

[101] W. Kuhn. Geospatial semantics: Why, of what, and how. *Journal on Data Semantics III (Special Issue on Semantics-based Geographical Information Systems)*, pages 1–24, 2005. (Cited on page 106.)

[102] R. Küsters. *Non-Standard Inferences in Description Logics*, volume 2100 of *Lecture Notes in Artificial Intelligence*. Springer, 2001. (Cited on pages 25, 59, 60, 98, and 101.)

[103] G. Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. The University of Chicago Press, Chicago, 1987. (Cited on pages 54 and 57.)

[104] L.B. Larkey and A.B. Markman. Processes of similarity judgment. *Cognitive Science*, 29(6):1061–1076, 2005. (Cited on page 8.)

[105] M. E. Lassaline. Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(3):754–770, 1996. (Cited on page 11.)

[106] A.-K. Lautenschütz, C. Davies, M. Raubal, A. Schwering, and E. Pederson. The influence of scale, context and spatial preposition in linguistic topology. In Thomas Barkowsky, Markus

Knauff, Gérard Ligozat, and Daniel R. Montello, editors, *Spatial Cognition V: Reasoning, Action, Interaction, International Conference Spatial Cognition 2006, Revised Selected Papers*, volume 4387 of *Lecture Notes in Computer Science*, pages 439–452, Bremen, Germany, September 2006. Springer. (Cited on page 77.)

[107] B. Li and F. Fonseca. Tdd - a comprehensive model for qualitative spatial similarity assessment. *Spatial Cognition and Computation*, 6(1):31–62, 2006. (Cited on page 8.)

[108] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998. (Cited on pages 57 and 58.)

[109] B. C. (2000) Love. A computational level theory of similarity. In *Proceedings of the Cognitive Science Society*, pages 316–321, Chicago, USA, August 2000. (Cited on page 11.)

[110] C. Lutz and R. Möller. Defined topological relations in description logics. In M.-C. Rousset, R. Brachman, F. Donini, E. Franconi, I. Horrocks, and A. Levy, editors, *Proceedings of the International Workshop on Description Logics*, pages 15–19, Gif sur Yvette (Paris), France, September 1997. Université Paris-Sud, Centre d'Orsay. (Cited on page 55.)

[111] M. Lutz and E. Klien. Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science*, 20(3):233–260, March 2006. ISSN 1365-8816. (Cited on pages 1, 13, 16, 29, 30, and 43.)

[112] D. Mallenby and B. Bennett. Applying spatial reasoning to topographical data with a grounded ontology. In F. Fonseca, A. M. Rodrígues, and S. Levashkin, editors, *GeoSpatial Semantics, proceedings of the Second International Conference (GeoS 2007)*, number 4853 in Lecture Notes in Computer Science, pages 210–227, Mexico City, November 2007. Springer. (Cited on page 96.)

[113] T. Mantay. Fbi-hh-m-291/2000: Commonality-based abox retrieval. Technical report, Department of Computer Science, University of Hamburg, Germany, 2000. (Cited on page 59.)

[114] T. Mantay and R. Möller. Content-based information retrieval by computation of least common subsumers in a probabilistic description logic. In *Proceedings of the International Workshop on Intelligent Information Integration, (ECAI'98)*, Brighton, UK, August 1998. (Cited on page 59.)

[115] E. Margolis and S. Laurence, editors. *Concepts: Core Readings*. The MIT Press, 1999. (Cited on pages 20, 43, 58, 98, and 100.)

[116] D. Mark, A. Turk, and D. Stea. Does the semantic similarity of geospatial entity types vary across languages and cultures? In *Workshop on Semantic Similarity Measurement and Geospatial Applications (at COSIT 2007)*, Melbourne, Australia, September 2007. (Cited on page 100.)

[117] A. B. Markman. *Similarity and Categorization*, chapter Structural alignment, similarity, and the internal structure of category representations., pages 109–130. Oxford University Press., Oxford, UK, 2001. (Cited on pages 11 and 15.)

[118] A. B. Markman and E. Dietrich. In defense of representation. *Cognitive Psychology*, 40:138–171, 2000. (Cited on page 73.)

[119] A. B. Markman and D. Gentner. Structural alignment during similarity comparisons. *Cognitive Psychology*, 25(4):431–467, 1993. (Cited on pages 10 and 11.)

[120] A. B. Markman and D. Gentner. Structure mapping in the comparison process. *American Journal of Psychology*, 113:501–538, 2000. (Cited on pages 10, 11, and 15.)

[121] A. B. Markman and C.H. Stilwell. Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(4):329–358, 2001. (Cited on pages 8 and 54.)

[122] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. WonderWeb deliverable D17. the WonderWeb library of foundational ontologies and the DOLCE ontology. Technical report, ISTC-CNR, 2002. (Cited on page 21.)

[123] D. Medin, R. Goldstone, and D. Gentner. Respects for similarity. *Psychological Review*, 100(2):254–278, 1993. (Cited on pages 1, 7, 11, 13, 15, 43, 57, 73, 78, and 100.)

[124] R. Molitor. Structural Subsumption for $\mathcal{ALN}$. LTCS-Report LTCS-98-03, LuFG Theoretical Computer Science, RWTH Aachen, Germany, 1998. (Cited on pages 39, 40, 65, and 98.)

[125] R. Möller. Expressive description logics: Foundations for practical applications, habilitation thesis. University of Hamburg, Computer Science Department, July 2001. (Cited on page 93.)

[126] R. Möller, V. Haarslev, and B. Neumann. Semantics-based information retrieval. In *IT&KNOWS-98: International Conference on Information Technology and Knowledge Systems*, pages 49–56, Budapest, Hungary, September 1998. (Cited on pages 1, 60, and 97.)

[127] D. R. Montello and P. C. Sutton. *An Introduction to Scientific Research Methods in Geography*. Sage Publications Ltd., 2006. ISBN 141290286X. (Cited on page 84.)

[128] D. Nardi and R. J. Brachman. *The description logic handbook: theory, implementation, and applications*, chapter An introduction to description logics, pages 1–40. Cambridge University Press, New York, NY, USA, 2003. ISBN 0-521-78176-0. (Cited on pages 25 and 96.)

[129] K. Nedas and M. Egenhofer. Spatial similarity queries with logical operators. In T. Hadzilacos, Y. Manolopoulos, J. Roddick, and Y. Theodoridis, editors, *SSTD '03 - Eighth International Symposium on Spatial and Temporal Databases*, volume 2750 of *Lecture Notes in*

Computer Science, pages 430–448. Santorini, Greece, 2003. (Cited on page 8.)

[130] R. M. Nosofsky. Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23(1):94–140, 1991. (Cited on pages 56 and 57.)

[131] N. Noy and M. Musen. Smart: Automated support for ontology merging and alignment. In *Twelth Workshop on Knowledge Acquisition, Modeling, and Management*, Banff, Canada, October 1999. (Cited on page 21.)

[132] C. K. Ogden and I. A. Richards. *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism*. Harcourt, Brace & Company, New York, 1923. (Cited on page 20.)

[133] H. Op de Beeck, J. Wagemans, and R. Vogels. Asymmetries in stimulus comparisons by monkey and man. *Current Biology*, 13: 1803–1808, 2003. (Cited on page 17.)

[134] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The Measurement of Meaning*. University of Illnois press, Urbana IL, USA, 1967. (Cited on pages 16 and 17.)

[135] U. Priss. Formal concept analysis in information science. *Annual Review of Information Science and Technology*, 40:521–543, 2006. (Cited on page 62.)

[136] F. Probst. *Semantic Reference Systems for Observations and Measurements*. Ph.D. Thesis, Institute for Geoinformatics, University of Münster, 2007. (Cited on page 23.)

[137] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19:17–30, 1989. (Cited on pages 1, 12, 46, 61, and 63.)

[138] 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, Torino, Italy, November 2004. (Cited on pages 7 and 17.)

[139] M. Raubal. Mappings for cognitive semantic interoperability. In F. Toppen and M. Painho, editors, *AGILE 2005 - 8th Conference on Geographic Information Science*, pages 291–296, Lisboa, Portugal, May 2005. (Cited on page 15.)

[140] C. Riedemann. Matching names and definitions of topological operators. In A. G. Cohn and D. M. Mark, editors, *Spatial Information Theory. Foundations of Geographic Information Science, International Conference (COSIT 2005)*, volume 3693 of *Lecture Notes in Computer Science*, pages 165–181, Ellicottville, NY, USA, September 2005. Springer. (Cited on page 77.)

[141] E. L. Rissland. Ai and similarity. *IEEE Intelligent Systems*, 21(3): 39–49, 2006. (Cited on page 1.)

[142] A. Rodríguez. *Assessing Semantic Similarity Among Spatial Entity Classes*. Ph.D. Thesis, Department of Spatial Infromation Science and Engineering University of Maine, 2000. (Cited on pages 1, 83, and 96.)

[143] 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. (Cited on pages 1, 7, 8, 9, 10, 13, 14, 17, 39, 45, 55, 57, 59, 60, 61, 63, 83, 84, 90, 97, and 100.)

[144] E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1978. (Cited on page 54.)

[145] S. Schade. Semantic reference systems accounting for uncertainty - a requirements analysis. In *Proceedings of of the 5th International Symposium on Spatial Data Quality*, Enschede, The Netherlands, June 2007. (Cited on page 99.)

[146] A. Schmidt. Ontology-based user context management: The challenges of dynamics and imperfection. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, OTM Confederated International Workshops and Posters, AWeSOMe, CAMS, COMINF, IS, KSinBIT, MIOS-CIAO, MONET, OnToContent, ORM, PerSys, OTM Academy Doctoral Consortium, RDDS, SWWS, and SeBGIS. Proceedings, Part I*, volume 4278 of *Lecture Notes in Computer Science*, pages 995–1011, Montpellier, France, October 2006. Springer. (Cited on pages 14 and 100.)

[147] A. Schwering. *Semantic Similarity Measurement including Spatial Relations for Semantic Information Retrieval of Geo-Spatial Data*. Ph.D. Thesis, Institute for Geoinformatics, University of Münster, 2007. (Cited on page 16.)

[148] A. Schwering. Approaches to semantic similarity measurement for geo-spatial data - a survey. *Transactions in GIS*, 12(1):5–29, 2008. (Cited on page 7.)

[149] A. Schwering and M. Raubal. Spatial relations for semantic similarity measurement. In J. Akoka, S. Liddle, I-Y. Song, M. Bertolotto, I. Comyn-Wattiau, W-J. vanden Heuvel, M. Kolp, J. Trujillo, C. Kop, and H. Mayr, editors, *Perspectives in Conceptual Modeling: ER 2005 Workshops CAOIS, BP-UML, CoMoGIS, eCOMO, and QoIS.*, volume 3770 of *Lecture Notes in Computer Science*, pages 259–269. Springer, Klagenfurt, Austria, October 2005. (Cited on page 7.)

[150] R. Schwitter. A controlled natural language for the semantic web. *Journal of Intelligent Systems*, 17(1-3):125–142, 2007. (Cited on page 86.)

[151] R. N. Shepard. Toward a universal law of generalization for psychological science. *Science*, pages 1317–1323, 1987. (Cited on page 1.)

[152] E. Sirin and B. Parsia. Pellet: An OWL DL reasoner. In Volker Haarslev and Ralf Möller, editors, *Proceedings of the 2004 International Workshop on Description Logics (DL2004)*, volume 104 of *CEUR Workshop Proceedings*, Whistler, British Columbia, Canada, June 2004. (Cited on page 69.)

[153] E. Smith. Concepts and induction. In M. Posner, editor, *Foundations of cognitive science*, pages 501–526. MIT Press, Cambridge, 1989. (Cited on page 58.)

[154] L. B. Smith. *Similarity and analogy*, chapter From global similarities to kinds of similarities: The construction of dimensions in development, pages 146–178. Cambridge University Press, 1989. (Cited on page 16.)

[155] J. F. Sowa. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA, USA, 2000. (Cited on pages 10, 19, and 20.)

[156] J. F. Sowa. Categorization in cognitive computer science. In H. Cohen and C. Lefebvre, editors, *Handbook of Categorization in Cognitive Science*, pages 141–163. Elsevier, 2006. (Cited on page 83.)

[157] J. F. Sowa and A. K. Majumdar. Analogical reasoning. In A. in de Moor, W. Lex, and B. Ganter, editors, *Conceptual Structures for Knowledge Creation and Communication, Proceedings of the International Conference on Computational Science (ICCS2003)*, number 2746 in Lecture Notes in Artificial Intelligence, pages 16–36, Berlin, 2003. Springer. (Cited on page 83.)

[158] J. Spencer-Smith and R. L. Goldstone. The dynamics of similarity. *Bulletin of the Japanese Cognitive Science Society*, 4:38–56, 1997. (Cited on page 43.)

[159] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. *Data Knowledge Engineering*, 25(1-2): 161–197, 1998. (Cited on page 19.)

[160] W. Sunna and I. Cruz. Using the agreementmaker to align ontologies for the oaei campaign 2007. In *The Second International Workshop on Ontology Matching, collocated with the 6th International Semantic Web Conference (ISWC2007)*, Busan, Korea., November 2007. (Cited on pages 2 and 8.)

[161] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, May 2005. ISBN 0321321367. (Cited on page 56.)

[162] 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, May 1994. Morgan Kaufmann. (Cited on pages 55 and 101.)

[163] De Silva V Tenenbaum, J. B. and J. C. Lanford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2322, 2000. (Cited on page 1.)

[164] A. Third, B. Bennett, and D. Mallenby. Architecture for a grounded ontology of geographic information. In F. Fonseca, M. A. Rodrígues, and S. Levashkin, editors, *SeoSpatial Semantics, proceedings of the second international conference*, number 4853 in Lecture Notes in Computer Science, pages 36–50, Mexico City, Mexico, November 2007. Springer. (Cited on page 96.)

[165] W. S. Torgerson. Multidimensional scaling of similarity. *Psychometrika*, 30:379–393, 1965. (Cited on page 1.)

[166] D. Tsarkov and I. Horrocks. Fact++ description logic reasoner: System description. In *Proceedings of the International Joint Conference on Automated Reasoning (IJCAR 2006)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 292–297. Springer, 2006. (Cited on page 69.)

[167] A.-Y. Turhan, T. Springer, and M. Berger. Pushing doors for modeling contexts with owl dl a case study. In *Proc. of the Fourth IEEE International Conference on Pervasive Computing and Communications. Workshop on Context Modelling and Reasoning (CoMoRea'06)*, pages 13–17, Pisa, Italy, March 2006. IEEE Computer Society. (Cited on page 13.)

[168] A. Tversky. Features of similarity. *Psychological Review*, 84(4): 327–352, 1977. (Cited on pages 1, 7, 8, 10, 15, 42, 43, 45, 56, 57, 90, and 91.)

[169] M. Uschold. Creating, integrating and maintaining local and global ontologies. In *14th European Conference on Artificial Intelligence (ECAI 2000)*, Berlin, Germany, August 2000. (Cited on pages 22 and 29.)

[170] M. van Assem, M. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga. A method for converting thesauri to rdf/owl. In *3rd International Semantic Web Conference (ISWC2004)*, Hiroshima, Japan, November 2004. (Cited on pages 32 and 33.)

[171] S. Wandelt and R. Möller. Scalability of owl reasoning: Role condensates. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops, OTM Confederated International Workshops and Posters, AWeSOMe, CAMS, OTM Academy Doctoral Consortium, MONET, OnToContent, ORM, PerSys, PPN, RDDS, SSWS, and SWWS 2007, Proceedings, Part II*, number 4806 in Lecture Notes in Computer Science, pages 1145–1154, Vilamoura, Portugal, November 2007. Springer. (Cited on page 93.)

[172] Xiao Hang Wang, Da Qing Zhang, Tao Gu, and Hung Keng Pung. Ontology based context modeling and reasoning using owl. In *Proceedings of Workshop on Context Modeling and Reasoning (CoMoRea '04), in conjunction with the 2nd IEEE International Conference on Pervasive Computing and Communications*, pages 18–22, Orlando, Florida, USA, March 2004. IEEE Computer Society. (Cited on pages 13 and 99.)

[173] P. Weinstein and W. P. Birmingham. Comparing concepts in differentiated ontologies. In *Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management (KAW'99)*, Banff, Alberta, Canada, October 1999. (Cited on pages 58 and 61.)

[174] A. Wierzbicka. *Semantics: Primes and Universals*. Oxford University Press, Oxford, New York, 1996. (Cited on page 54.)

[175] M. E. Winston, R. Chaffin, and D. Herrmann. A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444, 1987. (Cited on page 33.)

[176] W. Yeh and L.W Barsalou. The situated nature of concepts. *American Journal of Psychology*, 119:349–384, 2006. (Cited on page 8.)

[177] R. Zwick, E. Carlstein, and D. V. Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1(2):221–242, 1987. (Cited on page 37.)

COLOPHON

This thesis and the presented implementation were prepared using free and open source software; especially: GNU/Linux, the Kile Latex Editor, LaTeX as document preparation system, the Inkscape vector graphics editor, the R project (and RKWard) for statistical computing, Eclipse and OpenJDK for the implementation, and the Protégé editor for ontology engineering. The Latex template used for this thesis is an extended version of the template by André Miede.

VERSICHERUNG

Hiermit versichere ich, dass ich bisher noch keinen Promotionsversuch unternommen habe.

Münster, Juni 2008

Krzysztof Janowicz

Hiermit versichere ich, dass ich die vorgelegte Dissertation selbst und ohne unerlaubte Hilfe angefertigt, alle in Anspruch genommenen Quellen und Hilfsmittel in der Dissertation angegeben habe und die Dissertation nicht bereits anderweitig als Prüfungsarbeit vorgelegen hat.

Münster, Juni 2008

Krzysztof Janowicz

Hiermit erkläre ich, nicht wegen einer Straftat rechtskräftig verurteilt worden zu sein, zu der ich meine wissenschaftliche Qualifikation missbraucht habe.

Münster, Juni 2008

Krzysztof Janowicz