FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

# CoMetO: A Cognitive Design Methodology for Enhancing the Alignment Potential of Ontologies

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

**Doktorin der Technischen Wissenschaften**

by

**Alexandra Mazak**
Registration Number 8952186

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Mag.rer.nat. Dipl.-Ing. Dr.techn. Rudolf Freund

The dissertation has been reviewed by:

| | |
|---|---|
| (Ao.Univ.Prof. Mag.rer.nat. Dipl.-Ing. Dr.techn. Rudolf Freund) | (Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl) |

Wien, 22.02.2012

(Alexandra Mazak)

# Erklärung zur Verfassung der Arbeit

Alexandra Mazak
Lerchenfelderstrasse 124-126/1/14, 1080 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasserin)

# Acknowledgements

This thesis has been carried out at the Information & Software Engineering Group (ifs), within the department of Software Technology and Interactive Systems. I would like to thank Dr.techn. Monika Lanzenberger for her encouragement to do this research work and her constructive criticism in the course of this thesis and the conference contributions, as well as her engagement concerning my employment at the institute in the course of the FINCA[1] project.

I thank my first supervisor, Prof. Dr.techn. Rudi Freund, for his trust in my independent work style, as well as the scope he gave me for creative thinking. I enjoyed the freedom to develop my research work. I thank my second supervisor, Prof. Dr. Dieter Merkl, for his time, helpful suggestions, and his encouragement in the course of finalizing the thesis. In particular I extend a special "thank you" to Bernhard Schandl, who supported me throughout my working process. I am grateful for his advices and our enlightening discussions that inspired my work, and his proof reading in the final stage of the work. I would also like to thank fellow colleagues, in particular Manuel Wimmer for practical support, and Patrick Klaffenböck for fruitful discussions. My thanks also go to Edgar Weippl from Secure Business Austria (SBA) for the financial support, which gave me the opportunity to participate in the KEOD 2010 conference in Valencia (Spain), and the 3-month employment in the course of the FAMOS[2] project.

Finally, I would like to express my heartfelt gratitude to Harald for his love, understanding and encouragement since the very beginning. Without his support I would have never been able to go my way in studies and research works.

Alexandra Mazak
August $24^{th}$, 2011

[1] Female Intrapreneurship Career Academy, `http://www.finca.or.at/` (last accessed August-24-2011).

[2] Female Academy for Mentoring, Opportunities and Self-Development, `http://www.famos.or.at/` (last accessed August-24-2011).

For Henry and Harry

# Abstract

A multitude of research works and surveys dealing with *ontology alignment* point to an open research issue—there exists a gap between ontology engineers and users. That obstacle is caused by the discrepancy between the characteristics of the ontology's original modeler (e.g., their view of the domain, context-dependent background information, or modeling experience) and the characteristics of the user responsible for performing the alignment task (e.g., domain knowledge or prior exposure to ontologies). We assume that modelers can bridge this gap by starting to give information about their *expert knowledge* of the ontology's design process to users (e.g., for aiding meaning interpretation). For this purpose we introduce *CoMetO*—a cognitive design methodology—where we focus on the socio-technical component in ontology engineering, which involves contexts and perspectives. Our major aim is to foster an *evidence-based communication* from engineers to users.

In our approach the user support in ontology alignment becomes already important *ex ante* when an ontology's development process starts. We exploit expert knowledge of the original developers themselves; unlike other techniques which consider only *ex post* knowledge (e.g., derived from the ontology's structure). Our idea is to adapt a theory of pragmatics in order to supplement the ontology's relational structure with (context-based) *cognitive semantics* to provide—in combination with model-based semantics—a "complete package" for meaning interpretation as input in the alignment process. Therefore, we take a pragmatic-based point of view in CoMetO by considering the modelers' *cognitive perspective* on the domain as a component of their expert meta-knowledge. This subtle perspective is currently not reflected, neither in ontology engineering nor in ontology alignment. There exists no access to that knowledge. For this purpose we introduce a method by which the linkage of that perspective to schema level entities (e.g., classes and the binary relations that hold among them) is facilitated by *cognitive constraints*. In that process the modeler constrains the (intended) meaning of these entities in certain contexts.

Frequently, ontologies that describe the same domain of interest are similar but also have many differences, which are known as *heterogeneity*. There are various reasons for heterogeneity leading to different forms of it (e.g., pragmatic and structural heterogeneity). Aligning entities which are not meant to be used in the same context, or which follow different modeling conventions, may result in a mismatch. Users would benefit from knowing the *risk level of mismatch* between two sources prior to initiating an alignment. For this purpose we introduce two *mismatch-at-risk metrics* adapted from a risk metric of financial statistics and concepts of inferential statistics. The input of those metrics are automatically-computed, indicator-based metadata that are based on the cognitive constraints made by the modeler. The computed *mismatch-at-risk parameters* are predictors of a possible pragmatic-, as well as a structural heterogeneity caused mismatch. Our aim is to disburden users from a cognitively complex, time-, and cost-intensive task.



**Keywords:** Ontology design process, cognitive engineering, relevance theory, ontology alignment, structural and pragmatic heterogeneity, mismatch-at-risk metrics.

# Kurzfassung

Eine Vielzahl an Beiträgen und Studien zum Thema *Ontology Alignment* machen auf einen nach wie vor offenen Forschungsschwerpunkt aufmerksam—die Kluft zwischen Ontology Entwicklern und Usern. Dieses Hindernis resultiert aus den unterschiedlichen Charakteristiken dieser Personengruppen. Auf der einen Seite steht der Entwickler mit seiner Wahrnehmung des Domänen-Bereichs zum Zeitpunkt der Entwicklung der Ontologie, seinem Hintergrundwissen und seiner Modellierungserfahrung. Auf der anderen Seite steht der User mit seinem Wissen über den Domänen-Bereich und seiner Erfahrung im Umgang mit Ontologien. Wir gehen davon aus, dass die Entwickler diese Kluft überbrücken können, indem sie gezielt ihr Expertenwissen über den Ontologie Designprozess an den User übermitteln (z.B. als Hilfe für ein verbessertes Verständnis der Ontologie). Zu diesem Zweck haben wir *CoMetO* entwickelt. Unser Fokus ist dabei auf die sozio-technischen Komponenten im Ontology Engineering gerichtet, die ebenso Kontexte wie auch Perspektiven miteinschließt. Unser Hauptziel besteht darin eine Kommunikation vom Entwickler zum User zu ermöglichen, die auf Hinweisen aufgebaut ist.

In CoMetO berücksichtigen wir das Alignen einer Ontologie bereits vor der eigentlichen Ausführung, d.h. zum Zeitpunkt des Designs. Wir nutzen das Expertenwissen der Entwickler und unterscheiden uns damit von jenen Techniken, die Domänen bezogenes Hintergrundwissen lediglich im Nachhinein ableiten (z.B. aus der Struktur der Ontologie). Unser Ansatz besteht darin eine im Bereich der Pragmatik angewendete Theorie zu adaptieren, um die relationale Struktur einer Ontologie mit Kontext basierter kognitiver Semantik anzureichern, um—in Kombination mit Modell basierter Semantik—ein "Gesamtpaket" als Input für den Alignment-Prozess bereit zu stellen. Ziel ist eine verbesserte Interpretation von Bedeutungsinhalten. Aus diesem Grund nehmen wir eine Pragmatik orientierte Sicht in CoMetO ein, indem wir die *kognitive Perspektive* der Entwickler auf die Domäne als eine Komponente ihres Experten(meta)wissens berücksichtigen. Diese "feinsinnige" Perspektive wird gegenwärtig weder im Ontology Engineering noch im Ontology Alignment abgebildet. Es gibt keinen Zugang zu diesem Wissen. Zu diesem Zweck führen wir eine Methodik ein, die es dem Entwickler ermöglicht, diese Perspektive mit Schema-Entitäten (z.B. Klassen und den binäre Relationen zwischen diesen) zu verlinken. Dabei werden, ausgehend vom Entwickler, *kognitive Beschränkungen* dem Bedeutungsgehalt dieser Entitäten innerhalb bestimmter Kontexte auferlegt.

Häufig ähneln sich Ontologien, die den gleichen Interessensbereich beschreiben, sie weisen gleichzeitig aber auch viele Unterschiede auf. Diese werden allgemein als *Heterogenitäten* bezeichnet. Heterogenitäten haben verschiedene Ursachen, die zu unterschiedlichen Ausprägungen führen (z.B. pragmatische und strukturelle Heterogenität). Das Alignment von Entitäten, die nicht im gleichen Kontext verwendet wurden, oder die unterschiedlichen Modellierungskonventionen unterliegen, kann zu einem *Mismatch* führen. Für User wäre es von Nutzen, wenn sie Kenntnis vom Grad des Mismatch zwischen zwei Ontologien hätten und zwar bevor ein Alignment durchgeführt wird. Aus diesem Grund haben wir zwei *Mismatch-at-Risk Metriken* entwickelt, die wir von einer gängigen Risikometrik aus der Finanzstatistik und Konzepten der Inferenzstatistik abgewandelt haben. Zur Berechnung dieser Metriken verwenden wir indikatorbasierte Metadaten, die aus den kognitiven Beschränkungen des Entwicklers resultieren. Die berechneten *Mismatch-at-Risk-Parameter* sind Prädiktoren für einen möglichen Mismatch auf-

grund von pragmatischer wie auch struktureller Heterogenität. Die Zielsetzung ist, dem User einen kognitiv komplexen, zeit- sowie kostenintensiven Prozess zu ersparen.

**Schlagwörter:** Ontologie Designprozess, Cognitive Engineering, Relevanztheorie, Ontology Alignment, strukturelle und pragmatische Heterogenität, Mismatch-at-Risk Metriken.

# Contents

CHAPTER 1

# Introduction

John Godfrey Saxe's story about the "*Blind Men and the Elephant*" [Saxe, 1887] is seen as a metaphor in many disciplines. The story points out that different views of six men on parts of the same thing, namely an elephant, cause a conflict among these men and their perspectives. The analogy is used to demonstrate the need for the "big picture" of something to overcome diverging versions of reality. In the figurative sense this story addresses the problem of different world views of a domain in the development task of *ontology engineering*. It is a matter of common knowledge in ontology reuse that there exists no "global view" in modeling ontologies, even though they describe the same domain of interest, at the same level of detail. Applying Saxe's perspectival theory of the world to open issues in ontology reconciliation we can state that the major difficulties in evaluating and aligning ontologies are: (i) their domain or application dependency; (ii) the different purpose for developing ontologies, which cause different design goals; accordingly (iii) the modelers' diverging intentions on the usage of entities; (iv) different dimensions of context-dependent representations resulting from a variety of perspectives; and (v) the lack of providing cognitive aids as well as domain-related background knowledge to users for improving their decision-making process when aligning ontologies. These issues motivate us to implement an approach that addresses the problems caused by different world views in the context of ontologies describing the same domain of interest, and to analyze mismatch risks resulting from structural and pragmatic heterogeneity, which may occur when aligning ontologies.

## 1.1 Background

The vision of the Semantic Web is to provide a global infrastructure for the representation and exploitation of human knowledge. *Ontologies* are a central element of this vision providing the structural representation of that knowledge [Berners-Lee et al., 2001]. They enable parties to communicate and exchange knowledge. *Interoperability* among ontologies is the major goal for realizing the Semantic Web. Therefore, ontologies, as resources of the web, have to be reconcilable for gaining interoperability. The most well-established definition of an ontology is that

introduced by Gruber [1995]: "*an ontology is an explicit specification of a conceptualization*". This technical view of an ontology is often extended; just as Guarino [1998] explains in more detail: "*an ontology refers to an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words*". In short, an ontology is an artifact representing a scope of a real world domain, for a specific purpose, at a certain level of detail. It defines a common vocabulary for ontology authors (e.g., system analysts, domain experts, engineers), who need to share information of a domain. The ontology together with a set of instances constitute a *knowledge base* [Noy and McGuinness, 2001].

On a symbolic level, ontologies are logical theories with model-theoretic semantics expressed in an ontology language (e.g., F-Logic, LOOM, RDF(S), OWL), which provides the syntax and semantics. Thus, their expressiveness is language dependent. This means that the interpretation of an ontology is not left to the users, but it is explicitly defined by the semantics, which provides the rules (e.g., first-order logic) for interpreting the syntax [Euzenat and Shvaiko, 2007]. Therefore, the richer the semantics of a language is, the bigger is its expressiveness. There are:

- *top or upper level ontologies*, which describe general or domain-independent concepts (e.g., time, event, etc.);
- *task ontologies*, for describing generic types of tasks or activities;
- *core ontologies*, which provide global and extensible models into which data, originating from distinct sources, can be mapped and integrated;
- *application ontologies*, which describe domains of interest in an application-dependent manner; and
- *domain ontologies*, describing specific domains of the world.

However, most developed ontologies are placed under the concept of a domain ontology [Kalfoglou, 2000]. For detailed definitions about various kinds of ontologies we refer to Gómez-Pérez et al. [2003]. In this thesis we focus on domain ontologies, which are defined as: "*a set of definitions of terms that refer to a particular domain, together with some constraints on their use*" [Visser and Cui, 1998]. Thus, they are constructed to represent domain-relevant knowledge for a specific purpose.

Generally, domain ontologies are independently developed. Thus, they often expose differences in their structure, terminology, syntax, and semantics. That can obviously occur when ontologies describe different domains of interest. "*However, it also occurs even if they model the same real world domain, just because they were developed by different people in different real-world contexts*" [Rahm and Bernstein, 2001]. The differences between ontologies are known as *heterogeneity*, which is rooted in diversity in ontology modeling at different levels.

"*Ontology alignment is the process of discovering similarities between two source ontologies*" [de Bruijn et al., 2006]. Alignment is used to bridge heterogeneity in order to make ontologies and corresponding instance data interoperable. The goal is to find a balance between

heterogeneity and interoperability to make knowledge resources accessible. Ontology alignment is a cognitive complex, time-consuming task of ontology reconciliation. It makes ontologies consistent and coherent with one another keeping them separately (i.e., local) [Hameed et al., 2004]. The alignment is a set of pairs expressing the correspondences between the entities of two or more ontologies [Euzenat et al., 2005]. The correspondences are stored separately, and are therefore not part of the ontologies themselves.

Usually, ontology alignment is an iterative process. Firstly, an algorithm produces candidates of the selected ontologies to be mapped by a pair-wise similarity measuring. Secondly, the user examines these candidates and that information is given back to the algorithm, which produces more candidates. Thereby, the examination of candidates (e.g., accepting, rejecting, or changing) made by users is a critical step. In this specific task they should be supported to reduce their cognitive load of the candidates' verification.

Primarily introduced by Ehrig and Sure [2004], and presented in more detail by Ehrig [2007], an alignment process is made up of six main steps:

1. *feature engineering*, by which only parts of the ontology definition are selected to describe a specific entity (e.g., identifiers, labels, etc.);
2. *search step selection*, to determine a search space of candidate alignments in order to choose which entity pairs from the sources should be considered;
3. *similarity computation*, in this step the similarity values of candidate pairs are determined by using similarity functions (e.g., Levenshtein's *edit distance*[3] to compare string similarity);
4. *similarity aggregation*, there are several similarity values for a candidate pair of entities, which have to be aggregated in a single similarity value;
5. *interpretation*, generally a threshold is used to interpret the aggregated similarity values; and finally
6. *iteration*, which is terminating when no new alignments are proposed.

In a subsequent iteration one or several steps (e.g., $1-5$) may be skipped.

The goal of the alignment process is to find for each entity in one ontology a corresponding entity in the second ontology with the same or closest meaning for gaining *semantic interoperability*. This provides consensual understanding of the domain [Ehrig and Sure, 2004]. The major task in this process is the *mapping task*, which brings the entities of two or more ontologies into mutual agreement at a local level. In the literature the terms "mapping" and "alignment" are often used interchangeably. There exists no consensus about the usage of these terms. For instance: Ehrig [2007] describes the six steps as alignment process, whereas in a previous contribution [Ehrig and Staab, 2004] they denote these steps as mapping process. They point to mapping as also frequently called alignment. Euzenat and Shvaiko [2007] denote mapping as an oriented, or directed version of alignment (from one ontology onto another). We focus on the mapping task as an explorative, semi-automated task integrated in those steps where the user

[3]Edit distance, `http://nlp.stanford.edu/IR-book/html/htmledition/edit-distance-1.html` (last accessed October-19-2010).

is interacting with a tool, since the coherences among ontologies are often too complex to be automatically determined by the tools' algorithms. Therefore, the mapping task requires an efficient coordination between the user and the alignment tool. Falconer [2007] comments in this context that it is time "*to begin focusing on the user's needs during the mapping process*",—for instance, during the interpretation step (step five) in order to intervene when there are doubtable alignments (e.g., house/mouse identified by comparing the string similarity) in order to increase the quality of the proposed mappings. Improving the *effectiveness* (i.e., quality) of the user-guided process in ontology alignment is necessary, because of the high costs associated with a new implementation from scratch. The aim is to use existing ontologies for building a new one.

## 1.2 Problem Description

A small-scale pilot study with five participants was initiated by Smart and Engelbrecht [2008]. The aim of the study was to explore factors, which could determine why mismatches occur in the first place. The participants were asked to develop OWL[4] ontologies describing the same domain of interest, *a suicide bomb attack*, at the same level of detail. The results reveal that no two subjects settle on the same representational solution, even though they were all provided with the same stock of information (e.g., concepts, relationship) about that specific domain. This fact leads the initiators to the conclusion that there are more profound (i.e., not only terminological) differences among entities. These are resulting from "*the differential use of ontology modeling formalisms to express content*" [Smart and Engelbrecht, 2008]. This means that ontology authors can use the same ontology language for describing an identical domain of interest, nevertheless, there are subtle differences among the domain ontologies. The study's result brings forward that the cognitive state (i.e., mental state) of the ontology authors, rather than the semantics of the used language, causes these differences. The outcome of a previous study, analyzed by Bouquet et al. [2002], comes to similar conclusions. In this study two ontologies explicitly represent two contexts, but they overlap on a common part. The authors point out that, despite of this overlap, there is no guarantee for users in ontology alignment that the modelers' conceptualization of the common part is the same. The objections of the authors of both studies are confirmed by our own evaluation survey, the results of which we present in Chapter 7. There, the participants highly agree that the ontology authors' *cognitive perspective* (i.e., their intentional mental state) on the domain, at design time, has significant impact on the usage of ontology entities (e.g., classes, relations).

Heterogeneity lies in the origin of ontologies and cannot be avoided in distributed and open systems as the Semantic Web, or e-commerce. Different kinds of heterogeneity occur at different layers and levels; as discussed by Ehrig [2007], Ehrig et al. [2004], and summarized by Euzenat and Shvaiko [2007]. They are critical factors, which influence the quality and success of an alignment process. Heterogeneity among ontologies leads to *mismatch*, which forms an obstacle for interoperability (e.g., to integrate information across ontologies). Mismatch risks are causing expenditure of time and raising costs in the alignment of ontologies. There is still a lack of supporting tools and methods to cope with some special forms of it. Visser et al. [1997]

[4]OWL = Web Ontology Language, `http://www.w3.org/TR/owl-features/` (last accessed June-11-2011).

introduce two main categories of mismatch: *conceptualization mismatch* and *explication mismatch*. Both are based on Gruber's definition of an ontology. The authors differentiate between the conceptualization of the domain and the explication of that conceptualization. The first is made during the design process in which classes, instances, relations, functions, and axioms are distinguished in the domain (e.g., the process of ordering the classes as a hierarchy). The second requires an ontology language and explicates the description of ontology entities. The authors explain in detail conflicts which may arise among the entities of ontologies and classify those under one of the two major categories [Visser et al., 1997, Visser and Cui, 1998].

Mazak et al. [2010b] discuss that users are uncertain due to possible mismatch risks in ontology alignment. There is a multitude of heterogeneities which cause problems for *syntactic*, *semantic*, and *pragmatic* interoperability. Several of those heterogeneities are well understood in computer science. For instance: syntactic heterogeneity is generally solved by transducers, by which the semantics of expressions can be preserved [Bouquet et al., 2004]. Additionally, there are efficient techniques (e.g., SAT[5]-based techniques, DL[6]-based techniques) making semantic interoperability feasible by model theory. In this thesis we focus on possible pragmatic and structural differences that may exist among ontologies. Therefore, the other types of heterogeneity are at this point only outlined. For more comprehensive and detailed reviews of these differences we refer to: Bouquet et al. [2004], Chalupsky [2000], Euzenat [2000], Euzenat and Shvaiko [2007], Klein [2001], Ouksel and Sheth [1999], Tolk [2006], Visser et al. [1997], Visser and Cui [1998]; and Smart and Engelbrecht [2008].

### Structural Heterogeneity

The fact that engineers model the same domain differently causes heterogeneity among the structures of ontologies. Different modeling styles induce *structural mismatch* resulting from *explication mismatch* [Visser et al., 1997]. This kind of heterogeneity (Chalupsky [2000] describes it as dissimilarity in *modeling conventions*) results from differences in the way concepts are described by the ontology authors. Klein [2001] states that this mismatch is caused by "*the explicit choices of the modeler about the style of modeling*". He comments further on: "*[...] a distinction between two classes can be modeled using a qualifying attribute or by introducing a separate class*". Structural heterogeneity between two ontologies causes schema incompatibility, i.e., it hinders structural interoperability. Smart and Engelbrecht [2008] have observed that the experience of ontology engineers has a great impact on their modeling style. They indicate that experienced engineers tend to use the semantics (e.g., property restrictions, complex class descriptions) of an ontology language, whereas those engineers with relatively low experience tend to avoid using the full range of semantics. For instance, to describe that mechatronics is an interdisciplinary field combining mechanics, electronics, and informatics experienced developers would use the set operator `intersectionOf` of OWL DL[7];

```
<owl:Class rdf:ID="Mechatronics">
  <owl:intersectionOf rdf:parseType="Collection">
```

[5] SAT = propositional satisfiability

[6] DL = description logics

[7] OWL DL is the description logics-based subset of OWL Full.

```
    <owl:Class rdf:about="#Electronics" />
    <owl:Class rdf:about="#Informatics" />
    <owl:Class rdf:about="#Mechanics" />
  </owl:intersectionOf>
</owl:Class>
```

whereas an unversed engineer may describe that concept by creating a class *mechatronics* and by using the object property *consistsOf* with the `rdfs:range` sets *mechanics*, *electronics*, and *informatics*;

```
<owl:ObjectProperty rdf:ID="consitsOf">
        <rdfs:domain rdf:resource="#Mechatronics"/>
        <rdfs:range>
            <owl:Class>
                <owl:unionOf rdf:parseType="Collection">
                    <owl:Class rdf:about="#Electronics"/>
                    <owl:Class rdf:about="#Informatics"/>
                    <owl:Class rdf:about="#Mechanics"/>
                </owl:unionOf>
            </owl:Class>
        </rdfs:range>
    </owl:ObjectProperty>
```

Structural differences of domain representations cause significant problems especially for graph-based alignment tools (e.g., Anchor-PROMPT). Anchor-PROMPT [Noy and Musen, 2001] is a semi-automatic tool for graph-based alignment operations. The goal of the algorithm is to automatically find semantically similar terms. The tool considers ontologies as directed labeled graphs with classes as nodes and slots as links (i.e., arcs). The `WalkPaths`[8]-algorithm of Anchor-PROMPT searches for correlations among classes between two ontologies by parallel traversing paths in subgraphs. The algorithm analyzes the paths and determines those classes which frequently appear in similar positions on similar paths. The subgraphs have a certain length, and they are limited by initial points. The notion of these initial points is *anchors*. They are manually defined by the user or automatically determined by *lexical matching* methods. The length of the path is the number of edges in the path, predefined by the user. The algorithm is incrementing the similarity score between two nodes reached in the same position in the paths. In each step the similarity score, a coefficient, is aggregating cumulatively. A path follows the links (directed labeled edges) between classes (nodes) defined by hierarchical relations (`is-a` links) or by slots and their domain and ranges. Thereby, the `PathGenerator`[9]-algorithm of Anchor-PROMPT makes a kind of iterative Breadth First Search (BFS), where non-local context is taken into account. Additionally, the algorithm joins the classes linked by a subclass-superclass relation into *equivalence groups*. This means that the algorithm "plugs" as long as there exists an

[8]prompt source*.zip, WalkPaths.java, `http://protege.stanford.edu/plugins/prompt/` (last accessed March-2-2010).

[9]prompt source*.zip, PathGenerator.java, `http://protege.stanford.edu/plugins/prompt/` (last accessed March-2-2010).

`is-a` relation the associated class in an equivalence group till a slot frame is next in the path. The authors recognize that their tool is limited. They point out that "*the approach does not work well when the source ontologies are constructed differently*" [Noy and Musen, 2001]. In a course of lecture about PROMPT,—an algorithm and tool for automated ontology merging and alignment [Noy and Musen, 2000], the lecturer, Natalya Fridman Noy requires: "*we need to develop additional knowledge base measurements*"; because, "*knowledge-bases (i.e., ontologies) are rarely evaluated*" regarding their appropriateness for alignment. Many graph-based alignment tools do not work well if one of the source ontologies is a deep one with many inter-linked classes, and the other ontology is a shallow one, where the hierarchy has only a few levels.

Generally, engineers bring in their skills, preferences, and experience when designing ontologies. "*Which representation to choose is in most cases just a matter of taste or convention*" [Chalupsky, 2000]. Additionally, the importance of a concept in the domain of interest is determining such modeling conventions. Noy and McGuinness [2001] pose the question about the modeling style of an ontology on a concrete example of a *wine and food ontology*. They point out that the decision whether a class or an attribute should be modeled depends on the importance of the concept. They indicate that if a concept is important in the domain of interest, then the engineer should create a separate class for describing it. For instance: for the representation of the program of events at a conference it is important to distinguish between a working event and a social, or administrative event. These events have different properties, therefore they should be modeled as separate classes. However, heterogeneity at the ontology layer due to different modeling styles causes *structural mismatch*, which is difficult to resolve [Visser et al., 1997].

### Difference in Perspective and Pragmatic Heterogeneity

One reason for *conceptual heterogeneity* at the ontology layer, which is also called *semantic heterogeneity* [Euzenat, 2001], is the *difference in perspective* when modeling ontologies. Euzenat and Shvaiko [2007] address the problem of various perspectives by an example of maps from a *spatio-temporal* point of view. There exists no "global view" in modeling ontologies, not even do they describe the same region of the world, at the same level of detail. Quite contrary, there are many subjective (local) views causing a variety of perspectives. Benerecetti et al. [2001] describe three kinds of *perspective representations*: (1) *spatio-temporal*, (2) *logical*, and (3) *cognitive*. For instance: *spatial reasoning* can be performed by information visualization tools using some form of graphical representation languages; heterogeneity resulting from differences in the *logical perspective* on a domain is solvable by model theory (e.g., SAT-based techniques), whereas the existence of differences caused by the *cognitive perspective* can only be accepted, since there exists no access to this perspective [Benerecetti et al., 2001]. That makes it currently unresolvable. The other kinds of semantic heterogeneity are: *difference in coverage*, resulting from different partial views of the domain; and *difference in granularity*, which bases on differences in the approximation of partial descriptions [Bouquet et al., 2002].

*Pragmatic heterogeneity* [Bouquet et al., 2004], which is defined as *semiotic heterogeneity* by Euzenat and Shvaiko [2007] is related to *context*, which is mainly based on the ontology's purpose. This heterogeneity is caused by differences in the modelers' usage of entities in a certain (domain-related) context. That may lead to problems in the users' interpretation, if they

are not aware of the entities' "usage in context". For instance: there are two classes identified as syntactically equal by an algorithm based on their label similarity

$$(O_A) : \textit{Author}, (O_B) : \textit{author}$$

a user cannot automatically infer that the classes' meaning in use (i.e., context) is similar, too; or that

$$(O_A) : \textit{Contribution}, (O_B) : \textit{article}$$

are similar, because they are used synonymously in the same context (e.g., to describe authors and their publications). Ehrig et al. [2004] point to that fact; they state that "*similar entities are used in similar context*". The main problem is to define *usage patterns* for discovering such a (context-based) similarity in an efficient way [Stojanovic, 2005], since there is a strong relation between *pragmatic* and *context* [Janiesch, 2010]. The main problem is that such knowledge is implicitly encoded in the ontologies' structure, and therefore not exploitable for users when aligning the sources. Currently, this profound heterogeneity is not predictable or solvable; neither by model theory-based methods nor by semiotic-based methods, which exploit the theory of signs. Therefore, pragmatic heterogeneity is still a continuous problem in ontology alignment.

### Missing Cognitive Aids in Ontology Alignment

Klein [2001] specifies two aspects of practical problems when aligning ontologies:

1. "*it is difficult to find the terms that need to be aligned*";
2. "*the consequences of a specific mapping (unforeseen implications) are difficult to see previously*".

The understanding of relations among entities of different ontologies is a cognitively difficult task, added with the user's uncertainty about a possible mismatch risk resulting from heterogeneity factors among the sources. Often, this task is tedious, time-consuming, and error-prone [Euzenat and Shvaiko, 2007, Rahm and Bernstein, 2001], since it requires deep domain knowledge, which is not visible to users. A major problem in ontology alignment is the lack of aids for supporting the interplay between the user, the tools, and that process. Currently, users are supported in the interpretation task (step five) by a threshold, which is provided by alignment tools. Such a threshold is an evidence by which an alignment can be derived from the aggregated similarity values (step four). For instance, the threshold can be an absolute term (e.g., manually set by experts), or a value based on the highest found similarity value in the aggregation task [Euzenat and Shvaiko, 2007].

Gašević et al. [2006] address the issue of the theory behind knowledge representation with the field of *cognitive science*. Briefly defined, cognitive science studies the nature of human mind. Examples for mental states and processes are: "*thinking, reasoning, creating, remembering, language understanding and generation, visual and auditory perception, learning, consciousness, and emotions*" [Clark, 2001]. Assistance to such cognitive work can be called *cognitive support* [Walenstein, 2002]. One of the key issues of cognitive science is the study of

human thinking in terms of representational structures in mind, similar to computer-based data structures. Cognitive theorists propose that different mental representations constitute the basis of different knowledge representations [Hofstadter, 1995]. In order to be practical a representation technique needs a formal notation (e.g., first-order logic) for representing knowledge. Description logics (DL) are expressive enough to represent knowledge formally. Bouquet et al. [2002] argue that, "*knowledge is not simply a matter of accumulating "true sentences" about the world, but is also a matter of interpretation schemas, contexts, mental models, perspectives, which allow people to make sense of what they know*". This means that techniques which make semantic interoperability feasible are not efficient enough for communicating such characteristics as the modeler's cognitive perspective on the domain to users in ontology alignment (e.g., to provide *pragmatic interoperability*). The authors introduce a process of *meaning negotiation* to enhance the interpretation of ontologies. In this process mappings can be dynamically discovered by agents (e.g., users, tools) through communication, experience, trial and error. These tasks are abstracted as "meaning negotiation", which needs for its efficiency well-grounded, domain-related background knowledge [Bouquet et al., 2002].

In the past few years, researchers have developed many tools and techniques for creating ontology alignment; but, there has been done very little research to provide cognitive support for users in this field [Bontas, 2005, Falconer and Storey, 2007, Falconer et al., 2007]. One of a user's abilities is sight. Therefore, visualization techniques are a popular approach for such a support. These techniques help users to navigate ontologies in order to communicate relevant information, and to display ontologies from different perspectives. For instance, the two Protégé plug-ins: AlViz[10] and Jambalaya[11]. AlViz [Lanzenberger and Sampson, 2006] is a multi-view method, which supports the alignment of ontologies visually and aids the user's understanding of alignment results. Jambalaya is a tool using SHriMP[12], a domain-independent visualization technique, for enhancing users in browsing and exploring complex information spaces. Some visualization techniques lack contextual references since only parts of the ontologies can be viewed. For instance, it is not possible to filter the core concepts without additional domain-related background information. Users who are unfamiliar with the sources will probably get lost. However, most research has been spent on developing new algorithms for alignment tools (e.g., to gain better performance).

Falconer et al. [2007] conducted an online user survey to investigate: how users actually construct ontology mappings. The initiators gathered the feedback and came to the conclusion: "*we believe that at this point the biggest productivity gains in mapping tasks will come from better cognitive support rather than from an improvement of precision and recall in matching algorithms*" [Falconer et al., 2007]. Already, Norman [1993] indicates that, "*without external aids, memory, thought, and reasoning are all constrained*". In another study, conducted by [Falconer and Storey, 2007], the process of the user's decision-making in the mapping task was surveyed. At the end of this survey the initiators summed up: it would be a benefit to implement more user support in ontology alignment in order to gain a reduction of their cognitive

[10] AlViz, `http://alviz.sourceforge.net/` (last accessed January-19-2011).

[11] Jambalaya, `http://protegewiki.stanford.edu/wiki/Jambalaya` (last accessed December-8-2010).

[12] SHriMP = Simple Hierarchical Multi-Perspective, `http://www.thechiselgroup.org/shrimp` (last accessed December-8-2010).

load (e.g., in the interpretation task), which in turn would provide greater productivity gains. Generally, users rely on external artifacts to support their interpretation of meaning. For instance: WordNet[13] is an external resource, which contains synsets or sense. It can be used as *lexical entry* for concept interpretation. Synsets are structures containing sets of terms with synonymous meanings. Each synset has a gloss that defines the concept that it represents. It is often used as a source to gain background information about words and phrases expressed in natural language. Such resources provide auxiliary information for users, but some of these are less helpful without background knowledge of the sources.

Users often criticize that only simple mappings are automatically discovered by alignment tools. The rest is left to themselves with little, or no tool support. They state that the same occurs when performing semi-automatic methods, which are based on a heuristic search for candidates, where they are also burdened to validate the suggested mappings [Falconer et al., 2007]. The analysis of the observed participants made in the course of the surveys [Falconer and Storey, 2007, Falconer et al., 2007] reveal that the users' decision-making process in that task is often similar. Firstly, it relies on label similarity, and secondly, on the internal and external structure of ontologies. Generally, the respondents require a better support, for instance: in finding efficient starting points for the mapping task to reduce complexity, in identifying the most similar areas between the sources to explore more potential mappings, and in the limitation of the scope to focus on smaller chunks, e.g., by automatically validating higher priority candidates first. Additionally, we suggest that users should be supported to verify if the intended (context-based) meaning of the used terms are similar, too (cf. example of $O_A$ : *Author*, $O_B$ : *author*). Briefly summarized: the user's decision-making process is affected by the complexity of analyzing suggested mappings and the insufficient tool support for that task. Further difficulties and missed aids linked to ontology reconciliation are discussed by a manageable number of contributions which address the issues from a theoretical point of view, as: Benerecetti et al. [2001], Bontas [2005], Ernst et al. [2005], Falconer [2007], Falconer and Storey [2007], Falconer et al. [2006, 2007], Fetzer [2004], Giunchiglia et al. [2006], Guha et al. [2004], Janiesch [2010], Lanzenberger et al. [2008], Mazak et al. [2010a,b], Smart and Engelbrecht [2008], Walenstein [2002], Wand and Weber [2002].

## 1.3 Concluding Remarks

We highlighted three problems in ontology alignment which are still open research questions: (1) structural heterogeneity, (2) pragmatic heterogeneity, and (3) the lack of cognitive aids. In Section 1.2 we pointed to the phenomenon that ontologies which describe the same domain of interest are similar but also have differences, which are known as heterogeneity. This phenomenon is due to the fact that there exists no global view in modeling ontologies. We discussed that aligning entities which are not meant to be used in the same context, or which follow different modeling conventions, may result in unresolvable mismatches caused by pragmatic and structural heterogeneity.

[13] WordNet, a lexical database for English `http://wordnet.princeton.edu/` (last accessed January-13-2010).

We presented a multitude of contributions [Bontas, 2005, Noy and Musen, 2002, Smart and Engelbrecht, 2008] where the authors discuss that ontology engineers develop new ontologies from scratch, rather than (re)use existing ones. This corresponds to the conclusion that ontologies are rarely built to be reused, which is made by Simperl [2009] in the course of her feasibility study for reusing ontologies. Giunchiglia et al. [2006] put it in a nutshell: "*the lack of background knowledge, most often domain specific knowledge, is a hard one, and one of the key problems of matching systems these days*". However, for the purpose required in the summary of Falconer et al. [2007] and other cognitive support contributions [Bontas, 2005, Falconer and Storey, 2007, Giunchiglia et al., 2006, Janiesch, 2010, Mazak et al., 2010a, Smart and Engelbrecht, 2008], users need a method by which the modelers' intentions (i.e., what they have in mind when describing a domain of interest) are made visible and comprehensible when aligning ontologies. Therefore, *expert knowledge* of the original design process is needed for making such intentions transparent, which would favorably impact the interpretation step in order to gain more interoperability, in addition to the syntactic and semantic ones.

## 1.4 Objectives

Researchers have proposed many solutions to problems in ontology alignment. In the course of the literature research for our thesis we have found some interesting issues, which are still open tasks as summarized before (cf. Section 1.3). They provide the basis for the objectives in this thesis, which are:

1. to introduce a representation formalism for the modeler's cognitive perspective in order to make it visible to users when aligning ontologies;
2. to introduce a method by which the relevance of ontology entities can be evaluated based on their usage in certain contexts;
3. to generate additional indicator features for classes by which they can be ranked in lists in order to make their originally intended importance visible to users;
4. to provide predictors of potential structure- and pragmatic-based mismatch to users prior to starting an alignment process.

In the following chapters we explain our approach to meet these objectives and their influence on the main decisions which we made during work on the thesis.

CHAPTER 2

# State of the Art Integrating Contexts and Perspectives in Ontology Alignment

In this chapter we give an outline of current methods covering the same fields, as considered in our context- and perspective-based design methodology, in chronological order based on the publication date. This overview is not exhaustive. We refer to those contributions which are essential for the decisions made in our approach. Additionally, we discuss the stringent distinction between contexts and ontologies, which—as we assume—forms an obstacle for a fully supported meaning consideration as well as interpretation.

## 2.1 Outline of Alignment Techniques

Ontology alignment tools perform pair-wise comparison of entities from each of the source ontologies. Their algorithms are trying to find the best correspondences among these entities by selecting the most similar pairs. Ontologies are fairly complex structured. Thus, it is often practical to focus on different levels of ontologies separately, rather than trying to align those as a whole. The different methods for computing a similarity distance between the entities vary from *terminological*, *structural*, and *extensional* to *semantical* comparison [Euzenat and Valtchev, 2003].

*Terminological techniques* are based on the natural language. For instance: *extrinsic techniques* use external resources such as dictionaries or thesauri. The techniques determine similarity between lexical variations in the same term. They explore an equivalence between synonyms (e.g., *car* and *automobile*), and subsuming relationships between hyponyms (e.g., *car* and *motor vehicle*), but they do not cover the usage of classes.

*Structural techniques* are used to compare the internal and external structure of ontologies. They compute correspondences by analyzing how entities appear together in a structure. The

*internal structure comparison* exploits internal characteristics, as: cardinality, transitivity, or, symmetry of properties (e.g., attributes and relations). Frequently, it is possible to find multiple entities that represent similar internal characteristics, but, on closer examination it turns out that they are not similar. "*The internal structure does not provide much information on the entities to compare*" [Euzenat and Shvaiko, 2007]. Therefore, this method is commonly used in the initial alignment stage as a preprocessing step, or in combination with other techniques. *External structure techniques* treat ontologies as labeled graphs, with concepts as nodes and relations among those as arcs (directed edges). Such techniques analyze the position of nodes within the graph, viz. subgraph, of each ontology for similarity comparison. Context is derived from a node's neighborhood (i.e., arcs to other nodes). Generally, the techniques are based on the taxonomic structure, e.g., by counting the number of edges between the nodes. The results, which are given by these techniques, are not always semantically relevant [Euzenat and Shvaiko, 2007]. For instance: the class *Person* has more relations to other nodes in the hierarchy compared to the class *Author*, since *Person* is a superclass and *Author* is its subclass. Based on the number of edges it cannot be concluded that the meaning of these two classes is different, or that *Person* has more semantically relevance than *Author*. What is lacking is the reference to the domain context in which these two classes are used. In addition to its hierarchical structure, a graph contains a relational structure. Methods which explore this structure consider the concepts relations to other concepts. One major problem of this approach is that it is based on the entities usage, which is difficult to explore, because it is implicitly encoded in the structure.

*Extensional techniques* compare the instantiations of ontology classes. Knowing the classes' extensions provides information that is independent from the conceptual part of the ontology. The information is useful when a set of individuals, characterized in both ontologies, is available. This provides an easy way to compare the overlap between two classes. Problems may occur if such individual representations (instances) are not available, or if two ontologies do not share the same set of individuals.

*Model-* or *semantic-based techniques* are well-grounded, deductive methods which map elements according to their semantic interpretations. A deductive rule is a truth-preserving operation linked to the logical form of a sentence. For instance: SAT deciders are complete and correct decision procedures for propositional logics [Bouquet et al., 2003b]. Methods using SAT-based techniques for bridging the lack of missing background knowledge can only exploit unary predicates. Such techniques cannot handle binary predicates such as properties (e.g., `owl:DataTypeProperties`) or roles (e.g., `owl:ObjectProperties`). Modal-SAT can be used for extending the methods to binary predicates [Shvaiko and Euzenat, 2004]. The idea is to enhance propositional logics with modal logic operators. DL-based techniques can be used to establish relations between entities in a purely semantic manner. The advantage of using SAT-based techniques is that they support an exhaustive analysis of possible correspondences. However, "*pure semantic methods do not perform very well alone, they often need a preprocessing phase*" [Euzenat et al., 2004]; moreover, "*in the communication between people and computers, intelligibility cannot be ensured by semantics only*" [Euzenat, 2000].

Generally, alignment algorithms combine heuristic-based techniques on the basis of three criteria: (1) *syntactic*, (2) *semantic*, and (3) *structural* similarities among concept terms. Labels are the main distinguishing feature. Therefore, users consider those as a strong indicator for

similarity. The ontology structure helps in cases where labels do not work (e.g., when they are not expressive enough). Additional evidence may be provided by external resources such as dictionaries [Ehrig and Sure, 2005]. This fact corresponds to the results of the evaluated decision-making process of users in the task of ontology mapping, discussed by Falconer et al. [2007] (cf. Section 1.2). The evaluated data indicate that this process relies on the concepts' name similarity, followed by the structure (internal and external) of those concepts. The users feel confident about the correspondences if the structure of the candidates is similar. Therefore, most of ontology alignment approaches rely mainly on basic syntactical features of ontologies, as: the number of concepts and properties, their labels, and on algorithms taking into account the taxonomic structure (e.g., depth of an inheritance tree) of ontologies. These approaches are known as *syntax-driven techniques*. Semantics is not directly analyzed by such techniques. Mainly *element level techniques*, which are analyzing entities or instances of those entities in isolation (i.e., without considering their relations with other entities) use syntactical techniques [Giunchiglia and Shvaiko, 2003, Giunchiglia and Yatskevich, 2004]. Quite contrary, *semantic-driven techniques* want to map the *meanings* of concepts and not their labels, as in syntax-based approaches. Syntactic approaches are based on heuristics, which return similarity coefficients in the range of $[0,\ 1]$. Semantic techniques return logical relations (e.g., equivalence, subsumption) as output by exploiting model-theoretic information. This information is codified in the concepts and structures of ontologies. Further, detailed descriptions about existing approaches and tools are given for example by: Ehrig [2007], Ehrig and Sure [2005], Euzenat and Shvaiko [2007], Euzenat et al. [2004], Lanzenberger and Sampson [2006], Noy [2004], Rahm and Bernstein [2001].

## 2.2 Related Work

Considering *contexts and perspectives* in ontology reconciliation starts in the early $1990$'s with the work presented by Giunchiglia [1992]. He assumes that "*most cognitive processes are contextual in the sense that they depend on the environment, or context, inside which they are carried on*" [Giunchiglia, 1992]. He introduces a theory of reasoning with contexts, where he formalizes contexts as mathematical objects. His goal is to model reasoning as *deductive reasoning*, where a conclusion follows from a set of premises. An example for the method of deductive reasoning, which is commonly known, is the following:

| | |
|---|---|
| Premise (1): | all men are mortal. |
| Premise (2): | Socrates is a man. |
| Conclusion: | therefore, Socrates is mortal. |

The first premise states that all instances, which are classified as "men" have the attribute "mortal". The second premise states that "Socrates" is classified as a man, which is a member of the set "men". The conclusion states that "Socrates" must be mortal because he inherits this attribute from the classification as a man [Dethloff, 2001]. Giunchiglia structures the knowledge base as sets of facts $(A_1, \ldots,\ A_n)$. In his approach a context is a certain set of facts (e.g., $A_i$) used locally, e.g., to prove a given goal. He defines a context as "*the subset of the complete*

*state of an individual that is used for reasoning about a given goal*". He formalizes the reasoning method as a set of deductions, where each deduction is carried out inside a context. The resulting formal system is called *Multilanguage System (ML system)*, since each context has its own signature. For instance: the set of facts about an accounting system have "+" and "−" as parts of their signature, whereas the set of facts about authors and their publications have constants (e.g., *writes*, *Contribution*). A context can also contain facts which are abstractions of facts in another context. Each set of facts is associated with a language (e.g., PROLOG[14]) that is used to express context (e.g., in clausal form). In Giunchiglia's approach a context $c_i$ is defined as a triple $c_i = \langle L_i,\ A_i,\ \Delta_i \rangle$, where $\Delta_i$ is the set of inference rules associated with the set of facts $A_i$, which are written in $L_i$. Using contexts for formalizing the *locality* of reasoning distinguishes his approach from that proposed by McCarthy [1993], where context is introduced as a means for solving the problem of generality. Giunchiglia defines context as a world theory, which encodes an *individual's subjective view* of the world. On one hand, such a view is *partial* as the individual's description of the world is given by a set of contexts; and on the other hand, it is *approximate* as an individual never makes a description of the world in full detail. He assumes that there are different contexts which are theories of the same phenomenon (domain of interest) described at different levels of approximation. He focuses on a knowledge base as a set of interacting contexts $(c_1, \ldots,\ c_i, \ldots,\ c_n)$, where reasoning in one context may influence reasoning in the others. For this purpose he introduces a new set of rules by which a derived fact in one context can be linked to a fact derived in another context. He denotes such a linking rule as *bridge rule*. For instance: $A_j$ is derived in $c_j$ because $A_i$ is derived in $c_i$ by deductive reasoning, and the bridge rules link deductions in $c_i$ (which are reasoned first) to deductions in $c_j$. With his contribution he provides a basis for other research works considering context and perspectives in knowledge representation and reuse.

Ghidini and Giunchiglia [2001] present in their work a method which they call *Local Model Semantics* (LMS). They use LMS as a foundation for reasoning with contexts. They state that there are two principles underlying contextual reasoning: *locality* and *compatibility*. The intuition behind locality is that, "*reasoning only uses part of what is potentially available*". They denote such a part as *context of reasoning*, and, among such different parts there exists compatibility. In their approach a local set of models and a local domain of interpretation is used for mapping context. They differentiate between context as *partial object* and *complete object*. Formalizing context as partial defines it as a set of models, instead of the sense that each context is a single model (i.e., complete object). In this theory a context representation is a local model, described by a local language, with local semantics. This means that each context $c_i$ is associated with a certain formal language $l_i$ used to describe what is true in $c_i$. The semantics of $l_i$ is local to $c_i$ itself. Thus, each context has its own set of local models $M_i$, and local satisfiability relations $\models_i$. In their method they introduce *compatibility relations* and *domain relations*. The two relations are needed, because each specific local model is described by using a different first-order language ($l_i$). Therefore, each model is associated to a different interpretation domain. The compatibility relation between different representations is formalized by using domain relations, which relate the different interpretation domains. These relations can be perceived as logical constraints between different (logical) perspectives.

[14]PROLOG = PROgramming in LOGic.

Benerecetti et al. [2001] introduce general patterns of contextual reasoning, which base on the aforementioned works. They describe three theories of representation: (1) *partial*, which describes only a subset of the domain of interest; (2) *approximate*, which abstracts away some aspects of the domain that are not relevant for a given purpose; (3) *perspectival*, which encodes different kinds of perspectives on the domain. The perspecitval theory corresponds to the comment that "*an ontology describes a conceptualization, a view of a world from a particular perspective*" [Gruber, 1995]. There are three kinds of perspectives: (1) the *spatio-temporal view*, which considers the location and the point in time at which statements are used (e.g., using interval logic, or points for temporal representation); (2) the *logical view*, capturing the certain world in which the made assertions are true; and (3) the *cognitive view*, which encodes the modelers' intention,—their focus (e.g., based on beliefs, intentions, goals) when modeling the domain. The assumptions made in this approach are based on previous works: a conducted case study represented at AIMSA'98 [Benerecetti et al., 1998], and an article about their approach for contextual reasoning [Benerecetti et al., 2000]. The authors formalize context as the search for logical relations between the three forms of representation. Their approach is motivated by the theory of LMS and the framework of *MultiContext systems* (MCS) introduced by Giunchiglia and Serafini [1994]. In the latter, ML-systems [Giunchiglia, 1992] are presented from a technical point of view. These formal systems of multiple distinct logical languages can be used, as an alternative to modal logics, in the representation of physical perspectives. The authors divide a context dependent representation into three basic elements: (1) the *contextual dependencies*, which are a collection of parameters; (2) the *values* for each parameter; and (3) the *explicit representation*, which is a collection of linguistic expressions about the domain of interest. Additionally, they indicate that the mechanisms of contextual reasoning, which are studied in previous works by other researchers, generally fall into three abstract forms of contextual reasoning: (1) *expand/contract*, (2) *push/pop*, and (3) *shifting*. They state that each of these contextual reasoning operations consider one of the three forms of representation: partial, approximative, or perspective. They introduce these three basic patterns as the general patterns of contextual reasoning. The authors consider the sentences of an explicit representation (i.e., axioms of an ontology) inside a box and the related context outside that box. This metaphor of a box is adopted from the approach introduced by Giunchiglia and Bouquet [1997].

- *expand/contract*: they act on the assumption that an explicit representation, associated with a certain context, does not contain all the facts, but only a subset. Such a subset can be expanded to consider a larger collection of facts adjusted with a given goal, or a certain problem, and contrary acts the contract-operation. This reasoning mechanism allows to vary the degree of partiality.

- *push/pop*: the content of a context dependent representation is partly encoded in the sentences inside the box, and in the parameters outside the box. The push-operation produces an information flow from the inside to the outside, and vice versa the pop-operation. This mechanism allows to vary the degree of approximation.

- *shifting*: at least, certain contextual parameters can be changed without changing the whole collection. A changing of such parameters shifts the interpretation inside the box.

> For instance: the shifting-operation is related to the changing of the parameter time, viewpoints, which are depending on different positions, or categorizations.

They conclude, "*a logic of contextual reasoning is precisely a logic of the relationships between partial, approximate, and perspectival theories of the world*" [Benerecetti et al., 2001].

In the approach presented by Bouquet et al. [2002] ontologies are generally kept *local* to retain the *global knowledge* of ontology engineers (i.e., the identity of the community). In their framework semantic heterogeneity considers: *difference in granularity*, *difference in perspective*, and *difference in meaning* by using the same word meaning different things. They provide a language (CTXML[15]), which bases on XML[16] and XML Schema[17] for describing the *context space* of an ontology. This enables them to represent local ontologies as contexts, similar to the works of Bouquet and Serafini [2000], Ghidini and Giunchiglia [2001]. They take the three forms of representation introduced by Benerecetti et al. [2001]. Each local ontology represents a community's perspective on the domain of interest. There are possible relations between perspectives, which can be seen as mappings among such autonomous conceptualizations. The mappings represent directional relations between a context (source) and another context (target). They can be used to provide semantic-based services without destroying the semantic identities, which are inherent in each local ontology. The concept of context, akin to other works in this field, is an abstract representation. In their approach it contains: an *identifier*, additional *explicit assumptions*, which provide meta-information (e.g., the context owner, history), and, the *explicit representation*, which is the real content of a context represented as a labeled tree. The authors choose concept hierarchies as reference models. The concrete representation of a context is an XML document divided into two major parts: *header* and *content*. The header contains meta-information about: the owner of the context, the group which has developed the context, security information (e.g., access rights), and, history about how context was generated. They use DDLs[18], which is a KR[19]-based formalism, for providing *bridge rules* between the concepts at different abstraction levels.

In their subsequent work, Bouquet et al. [2003a] strictly hold on their distinction between ontologies and contexts. Ontologies are shared models, and contexts are local models that encode a party's subjective view of the domain (as firstly introduced by Giunchiglia [1992]). Their solution is to contextualize ontologies by keeping the content local, which means not shared with other ontologies. They create explicit *context mappings*, similar to the approach introduced by Ghidini and Giunchiglia [2001], by which contents can be mapped. They extend the syntax and semantics of the OWL language for representing contextual ontologies. They call this extension *Context OWL* (C-OWL). They point out that in other works "*several different ways of describing information semantics*" are used. They categorize these previous works in two broad approaches: (1) *ontologies*, which are shared models that encode a view common to a

[15]CTXML = ConTeXt Markup Language

[16]Extensible Markup Language, `http://www.w3.org/XML/` (last accessed January-11-2011).

[17]Extensible Markup Language Schema, `http://www.w3.org/XML/Schema` (last accessed January-11-2011).

[18]DDLs = Distributed Description Logics, `http://kedrigern.dcs.fmph.uniba.sk/reports/` (last accessed January-11-2011).

[19]KR = Knowledge Representation

set of different parties, and, (2) *contexts*, which are local models that only encode a subjective view of the domain. Therefore, ontologies are used for common representations (e.g., core ontologies), whereas contexts are used for autonomous representations with the need for a limited, controlled form of globalization. The authors admit that there are strengths and weaknesses of contexts and ontologies, in the way that one's strength is another's weakness. Ontologies make it possible to communicate between systems on a semantic level by defining a common understanding of specific terms. Their weakness is due to the fact that "*ontologies can be used only as long as consensus about their contents is reached*". Contexts are easy to define and maintain, but they encode no shared interpretation schemas. Their weakness is that explicit mappings among the elements of different contexts are required. They base their approach on two assumptions: if the parties are willing to share the intended meanings of their used terms this can be more easily supported by an ontology, and if ontologies contain information that should not be shared it is better to contextualize it. Therefore, in a *contextual ontology* the contents are kept local. They can be related with the contents of other (contextual) ontologies via explicit mappings. Such bridge rules allow to relate entities of different ontologies at the syntactic and semantic level. A set of bridge rules between two C-OWL ontologies is called *context mapping*. The constructs for representing bridge rules are taken from their previous work [Bouquet et al., 2002].

Bouquet et al. [2003b] view each semantic schema (e.g., concept hierarchies, ontologies) as context. In their contribution schemas are directed graphs whose nodes and arcs (directed edges) are labeled with terms from natural language. They consider only concept hierarchies in their approach. They define context as the partial and approximate representation of the world from a group's (e.g., ontology engineers, domain experts, etc.) perspective. This definition corresponds to that first made by Giunchiglia [1992], and extended by Benerecetti et al. [2001]. The authors state that "*a schema is the context in which facts are taken as true, decisions are made, objects are classified, relations among objects are asserted and understood*". They introduce an algorithm for automatically discovering relations across autonomous contexts which have well-defined semantics, and are directional. The implemented algorithm discovers bridge rules across contexts. They focus on the problem of discovering such semantic relations as a problem of *logical satisfiability* of a set of formulas. They point out that the meaning of a label depends on the context in which it occurs, and not only on the label's linguistic meaning. The algorithm is based on the concepts presented by Benerecetti et al. [2001], Giunchiglia and Bouquet [1997], where context is viewed as a box. The content of the box is a partial, approximate representation of the domain of interest. Contexts are mapped akin to the compatibility relations introduced in LMS [Ghidini and Giunchiglia, 2001]. The relation between two concepts of different contexts, discovered by the algorithm, can be seen as a *compatibility constraint* between the local models of the two concepts. The algorithm has two main phases: *semantic explication* and *semantic comparison*. The semantic explication makes the implicit information, which is hidden in the labels and structure of the concept hierarchy, explicit. Therefore, a logical formula is associated to each node of the conceptual graph, which encodes that information. For semantic comparison of two concepts, and their explicit encoded meaning, the problem of finding mappings between these concepts is transformed into a satisfiability problem. This can then be solved via SAT solver.

Magnini et al. [2003] introduce an algorithm by which semantic relations among concepts of different hierachical structures can be detected. The authors state: "*The semantics of schema models is not explicit but is hidden in their structures and labels*" [Magnini et al., 2003]. They want to make this information explicit to obtain semantic interoperability. They view ontology schemas as graphs. They focus only on the hierarchical structure of the graph to consider the hidden (semantic) information contained in that structure (i.e. the context). Besides the structural analysis of each source, their algorithm relies on a linguistic analysis. The concepts' labels are analyzed by using WordNet, which provides the labels' meaning. The authors consider an ontology's taxonomy as a natural language rooted tree. They derive the context in which a concept occurs from its label and position in the hierarchy. For instance: there is a node in the hierarchy with the label *Schools* with a descendant node labeled *US*. In this case the taxonomic relation between the two nodes has to be interpreted as a *location relation*. In their method each concept is analyzed separately as a stand alone object, and is associated with a logical formula of DL for interpretation. This makes it possible to associate a label of a node with a concept expression, a role description, or with an individual constant of DL. They define rules that should help to reconstruct a user's classification criteria. In the first rule ($M1$) it is stated that each concept has a meaning which is some entity of a world domain. In $M2$ they define that "*the* meaning *of a concept depends only on the labels associated with a finite set of nodes*", which they call the *focus of* $c$ ($F(c) \subset C$). This focus provides the information based on the position of $c$ in the hierarchical structure. "*Criterion* $M2$ *guarantees that the meaning of the concepts can be determined by visiting a finite (and possibly small) subset of the whole classification*" [Magnini et al., 2003]. On the basis of $M2$ the context in which $c$ occurs can be derived. Thus, the semantic meaning of a concept can be interpreted based on: the concept's label analyzed by using WordNet, and secondly, the ancestors of $c$ with their direct descendants (i.e., $F(c)$). Additionally, the authors apply standard classification criteria as used, e.g., in Yahoo![20]. For instance, in standard classification methodologies child nodes are always considered in the context of their parent nodes, and specialize the meaning of that nodes.

Guha et al. [2004] extend the aforementioned works presented by Bouquet et al. [2003a,b]. They represent a context mechanism for the Semantic Web. They review that "*we can no longer simply merge graphs without regard to where they occur*". Firstly, they point out that if the same data model and vocabulary is used subtle differences (e.g., pragmatic heterogeneity) between two representations may occur, which are resulting from the usage of terms at the task of conceptualization. This assumption corresponds to the comments made by Smart and Engelbrecht [2008], in the course of their pilot study's result-analysis (cf. Section 1.2). Secondly, they detect differences in the aggregation task. Ontology languages provide a method for aggregation at the data level (information layer). Higher level differences between knowledge representations make it sometimes inappropriate to directly merge data from the sources. Generally, assumptions are made for solving this problem, which lead to the use of same terms in different ways. The authors distinguish between the context mechanism developed in AI[21] and the requirements of context mechanism for the Semantic Web. They see the primary role for contexts on the Semantic Web to consider the differences between data sources when aggregating data of that

[20] `http://www.yahoo.com/` (last accessed January-13-2011).

[21] AI = Artificial Intelligence

sources. They define context as a resource, similar to the RDF[22] resource. They introduce a new concept—*AggregateContext*. In their approach they collect data from different URLs[23]. Each data source is abstracted into a context, which is defined as a *first class object* (similar to RDF resources). Each context is an instance of the class *Context*. They define a property type *contextURL* with its domain class *Context* in order to specify the location of the data source corresponding to the context to which it is abstracted. The content of the data source is assumed to be true in that context. More than one data source can be abstracted into a certain context. In a next step a context can be defined as an aggregation of data from other contexts, which makes it to an *AggregateContext*. That context is a subclass of the superclass *Context*. They denote the method for such an aggregation step as *lifting*. The difference between their approach and those where the method of *bridging* (i.e., bridge rules) form the basis to formalize contexts is that they use a technique by which content is imported and not linked. In the presented method the aggregation task is handled from a computational and a model-theoretic perspective. The presented approach is based on the authors' experiences which they have made in the course of a project named TAP[24].

Bontas [2005] takes the other line by introducing a *global context model* for ontologies. In her approach she describes usage patterns (syntactical, semantical, and pragmatical) to point out how contextual information may impact ontology reconciliation. She regards ontologies as an aid for a shared knowledge understanding, and a way to represent real world domains as domain ontologies. She states that to envision the Semantic Web, "*both ontology engineers and ontology users need a means to understand and evaluate existing ontologies*". Ontologies contain a valuable amount of knowledge, which cannot be easily evaluated by users regarding to the contexts in which they are modeled. In her work she illustrates that problem by a simple application scenario from the medical domain. The general approach is to improve reusability of available ontologies by providing a user-definable *application context* in order to enhance the development of more reusable ontologies. She presents critical factors which influence the ontologies' reusability by analyzing several real world use cases. She comes to the conclusion that currently ontology reuse is impracticable on a technological level: firstly, because of the different usage of formalization schemas, and secondly, because of the limitations of established tools to give users more information (i.e., intrinsic features) about the source ontologies themselves. She states that the absence of contextual information is a major obstacle to a wide-spread dissemination of ontologies. In the analysis of the case studies she finds out that usage-related information (i.e., context information) may improve several stages of a reconciliation process among ontologies. For this purpose she introduces a model for a formal declarative description of ontology-centered context information. She develops a *core context model* that can be used to specify context information in a transparent manner. Context information is formally defined as a meta ontology by using the advantages of the OWL vocabulary.

Giunchiglia et al. [2006] introduce a fully automated method to address the problems caused by the lack of background knowledge. Their approach is to use semantic matching iteratively.

[22]RDF = Resource Description Framework, `http://www.w3.org/RDF/` (last accessed January-13-2011).

[23]URL = Uniform Resource Locator.

[24]The Alpiri Project, `http://www-ksl.stanford.edu/projects/TAP/` (last accessed January-9-2011).

The two key ideas are: to compute logical relations among the entities of the sources instead of coefficients and to determine these relations by analyzing the meaning of that entities. For this purpose, the entity labels are translated into propositional formulas. This makes it feasible to transform the mapping problem into a propositional unsatisfiability problem, which is then resolvable by using SAT deciders (similar to Bouquet et al. [2003b]). In their method external resources, such as dictionaries, are used to fill the gap of missed background knowledge. They reference to other strategies, which may attack this problem, for instance: previous match results, manually declarations of missing axioms, upper level ontologies, and sense-based approaches as WordNet for defining semantic relations.

Falconer et al. [2006] present *CogZ*[25], a system for *Cognitive Support and Visualization for Human-Guided Mapping*. CogZ is a user-interface plug-in for the ontology management PROMPT-suite [Noy and Musen, 2003] implemented in Protégé[26]. The latest version bases on the requirements of cognitive aids for users in the mapping process, analyzed by Falconer and Storey [2007] (cf. Section 1.2). They choose Treemaps[27] for visualization, which scale well even if ontologies are large with thousands of nodes. Candidate-heavy regions are identifiable by color intensity. The *pie chart view* gives a detailed overview of the number of candidate mappings. Context is presented by the neighborhood of the mapping terms. The generated context provides a visual, structural comparison between two candidates. The number of mappings, explored by the PROMPT tool, can be reduced by filters (e.g., hierarchical filters). That makes it feasible to minimize the mapping scope. Thus, users can focus on certain mapping types. For instance: the candidate lists of PROMPT can be extended by temporary mappings, which are highlighted for supporting the user's working memory. Additionally, users can make annotations for explaining the chosen mapping of two terms. Semantic zooming is supported to highlight the user's current focus. The ontology trees can be filtered to display terms with, or without mappings. This can be done automatically when a user types in a search query.

Wagelaar [2008] addresses the problem that configuration constraints which are part of a configuration language (e.g., XML) are limited due to the language's expressiveness. To overcome this limitation he implements in his shortly introduced approach of *contextual constraints* a *context vocabulary ontology*, which can be separately used to describe constraints for a configuration language based on context. Such a separate formalism for context configuration makes it feasible to determine which contextual constraints are satisfied by a certain context. He uses the language OWL DL and its features to express contextual constraints as OWL classes. The `subClassOf` relation makes it feasible to structure these constraints in a hierarchy. He uses such a class hierarchy of context constraints in order to determine which configuration choice is more or less specific to context (e.g., the more specific the closer is the match to context). He uses the Eclipse Modeling Framework (EMF) [Budinsky et al., 2004] as a "superstructure" in order to bring the explicit defined context model into agreement with the metamodel of a configuration language. He points to the drawbacks of his approach as follows: the granularity is limited to constraints on schema level elements, thus the instances of a class must introduce the

[25] `http://www.stanford.edu/~sfalc/cogz/cogz.html` (last accessed January-14-2011).

[26] `http://protege.stanford.edu/` (last accessed January-15-2011).

[27] Treemap is a space-constrained visualization of hierarchical structures, `http://www.cs.umd.edu/hcil/treemap/` (last accessed January-14-2011).

same contextual constraints, and only classes can be context-constrained, but not their attributes.

Wu et al. [2008] present *CARRank*, a *Concept-And-Relation-Ranking*. CARRank is a flexible algorithm for identifying and evaluating the importance of concepts and relations in an ontology. The algorithm is flexible in that it can be easily applied to any RDF-based ontology. Additionally, it requires no user interaction. They state that their approach enhances the users' understanding of the sources. For instance, an interesting sub-scope of an ontology can be determined in order to take out parts for further computation. The authors view ontologies as directed labeled graphs. They use structural information of the graph in order to deduce the importance of concepts and relations. In their approach they reconstruct an ontology's design process by making assumptions of the domain and the ontology's design process. They introduce two features for this purpose: (1) a concept is more important the more relations are starting from that concept; (2) a concept is more important the more relations it has to other already important concepts. The authors state that these features might be the ones ontology engineers also would suggest to users for getting familiar with the ontology. The implemented algorithm weights relations (i.e., directed edges) in an iterative manner. Thereby, the importance of concepts and the weights of relations reinforce one another. This means that a concept is more important the higher the relation weight is to other concepts, and a relation weight is higher if it starts from a more important concept. The authors indicate that such concepts are good starting points which have the most relations to other concepts. They consider no prior knowledge (e.g., design process knowledge of the original developers) in their approach. The CARRank algorithm is akin to link analysis ranking algorithms on Web pages (e.g., tracking the user's browsing activities). They do not consider any context or perspective in their work. We refer to their approach because of the similar idea to rank concepts by their relevance to other concepts, as introduced in our approach (cf. Section 6.4).

Finally, we want to make a brief summary of the work of Janiesch [2010]. He states that artifacts as: ontologies, models, and methods are intended to solve problems. "*This entails that any situated use thereof can only function properly when embedded in its socio-technical context.*" In his context-based approach he proposes to regard the more general context of deployment (e.g., design task) at the model layer, rather than focusing on situational, ad hoc, details at the data layer. He points out that such a contextualization depends on the purpose-specifics of the domain. He underpins his assumptions with a quantitative analysis of journal articles. He views a representation, similar to the aforementioned approaches, as a partial, approximate abstraction from the original, but does so in compliance with *pragmatic* requirements. He summarizes that an artifact (e.g., ontology) is the representation of a domain of interest for the ends of a subjective (i.e., specific purpose), which is commonly based on a semi-formal language. He points out, analogous to Bontas [2005], Bouquet et al. [2002], Wand and Weber [2002], that a model should support the communication between developers and users. His justification to include context in the conceptualization task is based on the socio-technical design approach to *information systems development* (ISD). He proposes that conceptual modeling contains three parts: "*model, method, and their context as the pragmatic representation of social practice*". In his assumption he represents context as environment in which a model, or components of it (e.g., concepts, relations) have a certain meaning, according to van Dijk [1982], who comments: "*context is a theoretical construct necessary to interpret meaningful expressions of discourse*".

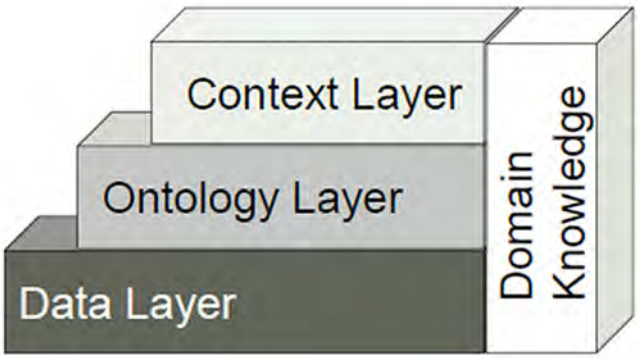



Figure 2.1: Layer model introduced by Ehrig et al. [2004].

At the end of his contribution he summarizes: "*it can be concluded that incorporating some concept of context in conceptual modelling methods is beneficial for the understanding of their content*".

## 2.3 Discussion

In all works described before the authors state that *contexts and perspectives* become more and more central in theories of knowledge representation and that they are both worth being considered in ontology reconciliation (e.g., ontology alignment). The authors [Benerecetti et al., 2001, Bouquet et al., 2002, 2003b, Ghidini and Giunchiglia, 2001, Giunchiglia, 1992, Guha et al., 2004, Magnini et al., 2003] stick to a stringent partition between ontologies and contexts. This partition is also reflected in the *layer model* shown in Figure 2.1. Ehrig et al. [2004] present in their contribution a framework for similarity measures among ontology entities at different layers. They distinguish between *data*, *ontology*, and *context layer*. At the bottom layer (data layer) entities are compared by considering the data types (e.g., string, integer) of their data values. For instance: strings can be compared by using generic similarity functions such as the edit distance. At the middle layer (ontology layer) the semantic relations among entities are compared, for instance, by using the graph structure of ontologies (e.g., taxonomy) for determining similarity. At the top layer (context layer) the entities' usage is considered in some external context. The context information is used external (separated) to the ontology. These layers are horizontally arranged, one upon the other. An additional layer, the *domain knowledge layer*, is vertically arranged. This orthogonal dimension, by which domain-specific aspects are considered, affects the other layers. At this layer auxiliary information of external resources (e.g., core ontologies, oracles) are often used for assessing the similarity among entities, since corresponding *background information* should be used for a more precise similarity computation. Ehrig [2007] continues the work from 2004 [Ehrig et al., 2004] by relating certain forms of heterogeneity to the layers (cf. Section 1.2). Besides, there are other works [Maedche and Staab, 2002, Zanobini, 2006] considering a layer architecture that are based on the same stringent partition.

Generally, ontologies are shared models and contexts are local models that encode the mod-

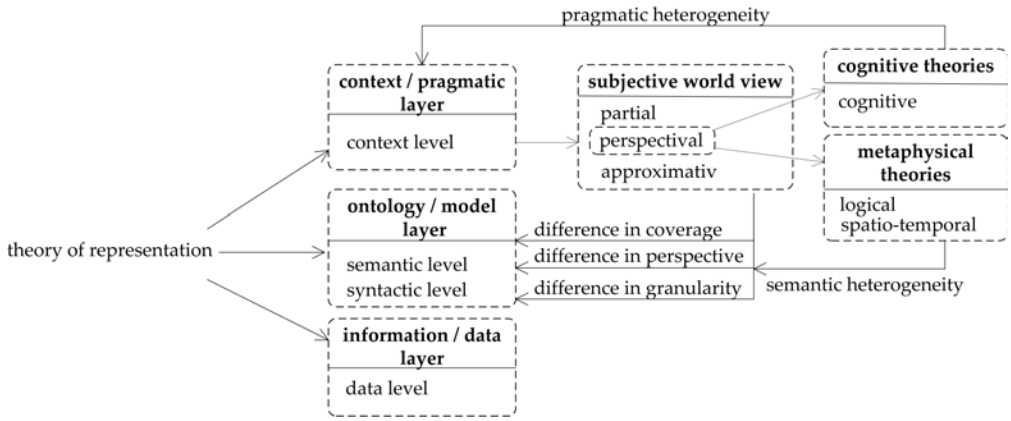



Figure 2.2: Overview of context, perspectives, layer classification, and heterogeneities.

eler's subjective view of the domain [Bouquet et al., 2003a]. First of all, associated with ontology reuse, Giunchiglia [1992] detects that context encodes an individual's subjective view of the world, which is partial and approximate. Benerecetti et al. [2001] associate to each of these views a certain definition (cf. Sections 1.2, 2.2) and specify a "purely" subjective view of the domain as *cognitive perspective*. They state that this subtle form of perspective "*is very important in the analysis of what is generally called an intensional context*" [Benerecetti et al., 2001].

There is a discrepancy that comes to mind when reviewing the multitude on definitions of perspectives and representation, the aforementioned partition, and the semantic- as well as pragmatic-based heterogeneity problem caused by differences in perspectives, which we introduced in Section 1.2. Figure 2.2 helps to illustrate this discrepancy. It gives an overview of the theories of contexts and perspectives, as well as the associated heterogeneity types. On one hand, there is a conceptual or semantic heterogeneity resulting from difference in perspective when describing a domain of interest. This heterogeneity is related to the *ontology layer*, which means that perspectives are related to that layer. On the other hand, there is a pragmatic heterogeneity caused by differences in the entities usage in context when modeling an ontology from scratch that is related to the *context layer*. Such heterogeneity may occur when the modelers' intended meaning on entities, which is mainly based on the purpose of the domain to be modeled differs (e.g., due to certain business goals). Obviously, there is a strong relation between the *cognitive perspective*, *pragmatic*, and *context*. The discrepancy is in that; there are perspectives (e.g., spatio-temporal) where differences of which are related to the ontology layer, and there is this strong relation between the cognitive perspective and the entities usage in a certain context. However, the existing separation of contexts and ontologies leads to a "separation of meaning", which constitutes an obstacle when providing users a single, consistent environment. Generally, semantics is expressed as *context-independent meaning* (sentence meaning), whereas pragmatics or cognitive semantics as *context-dependent meaning*. We argue, similar to Janiesch [2010], that it is not beneficial to relate meaning only to the ontology layer. This results in "cutting" any relation to context. We take a more *holistic view* in this field. Thus, we do not facilitate such a stringent differentiation between semantics and contexts as discussed. Fetzer [2004] underpins

our position. She points out that "*meaning is at the heart of both semantics and pragmatics*"; as a result of which we propose to consider pragmatics similarly to semantics at the ontology layer in order to fully support meaning consideration (in the design process), as well as meaning interpretation (in the alignment process).

We assume that the partition's beginning is to be found in the *theory of contexts* introduced by Bouquet and Serafini [2000]. They divide this theory into two major categories: *metaphysical theories* and *cognitive theories*. In metaphysical theories "*contexts are thought of as part of the structure of a world*". In cognitive theories "*contexts are part of the structure of an agent's cognitive state*". In the metaphysical approach context is "objective" and can be shared, whereas, in the cognitive approach the context is "subjective" and local. The knowledge of the subjective or *intentional context* may offer a much broader view of a domain, since it helps to reflect an engineer's view of the domain when describing it. A fact that researchers [Bontas, 2005, Bouquet et al., 2002, Falconer and Storey, 2007, Falconer et al., 2007, Rahm and Bernstein, 2001] are aware of it, since they propagate to implement a method by which access is given to such (design process) knowledge. They state that meaning negotiation from ontology engineers to users is made feasible by making this knowledge exploitable in order to move closer to the Semantic Web vision.

The introduced methods for deriving context (e.g., by bridge rules) are based on model theory. Therefore, these approaches can be subsumed under the metaphysical theory where context is a standalone object separated from an ontology. In the course of our literature research we found no approach in which both context and ontology are considered as a single environment. Janiesch [2010] proposes to equip the model layer with context in compliance with pragmatic requirements. However, he introduces neither a theoretical concept nor a method by which such a consideration can be made feasible in practice. The implementation of such a method would facilitate that contextualization can be made *ex ante* at the design process included in the description of domain concepts, which would make context information explicit (e.g., visible) to users.

## 2.4 Concluding Remarks

In Section 2.1 we gave an overview of the techniques of alignment methods. Mainly, these methods are geared to the ontology's syntactical features (e.g., number of classes), the labels of classes, semantic features (e.g., declarative formalism), and the taxonomical structure. All the discussed methods can only exploit *ex post* knowledge, which means knowledge that is derived from these features afterwards (at time of alignment) when ontologies are completed. That is why *meaning* can only be reconstructed without "well-grounded" intelligibility of the original engineering methodology, as noted by Bontas [2005] and Janiesch [2010]. This lack corresponds to the conclusion made by Biggerstaff and Richter [1987] that from the viewpoint of re-usability, the reuse of components at design stage has more potential for success than the reuse of code (e.g., formalism). Even if model theory-based techniques are used, as described by Bouquet et al. [2003b] (e.g., by associating logical formulas to concepts) the lack is that such a method can only approximate human-based interpretation. Additionally, the encoding of semantic mappings into logical relations, based on heuristics, may cause information loss.

We discussed that there exists no unifying theory of context, just as there exists no unified view of ontology engineers even though they describe the same domain of interest. Starting with the work presented by Giunchiglia [1992] and subsequent approaches introduced by Bouquet et al. [2002, 2003a], Ghidini and Giunchiglia [2001] compatibility relations are used, as *bridge rules*, to formalize contexts. Guha et al. [2004] introduce *lifting rules* to aggregate context from other contexts. They all consider context as local, and as separated from the ontology layer. In the recent past, Bontas [2005] argues that the associations between knowledge (contained in the ontology) and context are not easily detectable for users. Additionally, we illustrated that such a stringent partition leads to a discrepancy. We pointed to a theory where context is part of the modeler's mental state or *cognitive perspective* at design time. We identified that this intentional type of perspectival representation correlates with the pragmatic requirements proposed by Janiesch [2010].

From our analysis we can observe that the deep and unresolved problems in the field of ontology development and alignment are: (i) to link context information to the ontology layer at design time; (ii) to implement a method by which the modeler's cognitive perspective on the domain in various contexts is made explicit in order to improve meaning interpretation; and (iii) to implement a method by which problems (e.g., pragmatic heterogeneity) caused by different cognitive perspectives on the same domain are made visible to users in ontology alignment. This chapter does not attempt to provide an overall review of the state of the art in ontology alignment. We summarized those works, methods, and techniques, which are relevant for the assumptions made in our approach. Further detailed insights in the comprehensive research field of ontology reconciliation are presented by excellent and thorough contributions: Bouquet et al. [2004], Ehrig [2007], Euzenat and Shvaiko [2007], Euzenat et al. [2004, 2005, 2006], Hitzler et al. [2006], Rahm and Bernstein [2001].

CHAPTER 3

# Ontology Engineering

*Ontology engineering* is defined as "*the set of activities that concern the ontology development process, the ontology life cycle, and the methodologies, tools, and languages for building ontologies*" [Gómez-Pérez et al., 2003]. Ontology engineering contains three main research fields: (1) *ontology management*, (2) *ontology development*, and (3) *ontology support*. In this chapter we discuss the development of ontologies and its activities. We will not cover the other categories; for this purpose we refer to: Gandon [2002], Gašević et al. [2006], Gómez-Pérez et al. [2003], Gruber [1995]. We start with an outline based on the ontology development guide presented by Noy and McGuinness [2001]. We review, based on our discussion made in Section 2.3, that each engineering group has its own conceptualization of the domain of interest, which is partial, approximate, depending on cognitive aspects, and relevant concerning the fulfillment of the group's objectives. We focus on the social context of ontology development. For this purpose we present our view of the design task by comment on the rich *process knowledge* of the parties involved in that task. We introduce two example ontologies which both describe the same domain of interest in OWL DL. We use these ontologies in order to reveal the differences (e.g., in the ontologies' structure) resulting from the engineers' independent design decisions and different modeling styles.

## 3.1 Ontology Development

The ontology is an instrument that can be used to represent a domain in a structured way [Euzenat and Shvaiko, 2007]. The idea is that it provides constructs for users to organize information as taxonomies of concepts, each with their attributes, and to describe relations among concepts in order to represent that concepts' relationship in the real world domain. There are three phases of ontology development: (1) *pre-development*, (2) *development*, and (3) *post-development* [Gómez-Pérez et al., 2003]. In the first phase the environment and feasibility studies are made. The second phase contains: specification, conceptualization, formalization, and implementation. Finally, the third phase concerns maintenance and use/reuse of ontologies.

In our approach we focus on the process of creating domain ontologies from scratch, which is assignable to the second phase, and the use/reuse phase in schema-based ontology alignment.

Noy and McGuinness [2001] present an ontology development guide on the basis of a wine and food ontology. They use OWL as ontology language, and an early version of Protégé[28] as ontology editor. This guide is following other works: Gruber [1993], Grüninger and Fox [1995], Uschold and Grüninger [1996]. The main stages of ontology design are:

1. *Goal of the ontology:* The purpose of the domain represents the goals to be satisfied by the design process [Ramesh and Dhar, 1992]. Grüninger and Fox [1995] state in this context: "*we must agree on the purpose and ultimate use of our ontologies*".

2. *Determine the scope:* Ontology development starts by determining the domain and scope. The scope is *partial* and *approximate*. It is not the task of ontology design to describe all state of affairs of the domain. Therefore, representations are partial covering only domain knowledge that is believed to be relevant to the task at hand. Additionally, there is a certain level of granularity. A representation is approximate, because it abstracts away details that are not relevant for the ontology's purpose (cf. Section 2.2).

3. *Consider reuse of existing ontologies:* There are libraries of reusable ontologies on the Web (e.g., DAML[29] ontology library). They contain a judge amount of ontologies, which can be imported into the ontology development environment. If necessary, they can be translated from one formalism to another.

4. *Enumerate terms:* A useful way for the ontology authors is to write down a list of all the terms in the domain of interest which should be mapped in the model, as well as the relations among those terms. The main goal of ontology design should be that the abstract model reflects the objects and their behavior in the world domain, as close to reality as possible.

5. *Define classes and the class hierarchy:* There are several methods for developing a taxonomy: *top-down* starts with the definition of the most general domain concepts; *bottom-up* starts with the definition of the most specific concepts; and a *combination* of both approaches. "*The approach to take depends strongly on the personal view of the domain*" [Noy and McGuinness, 2001].

6. *Define the properties of classes:* Classes alone provide not enough information; according to Passin [2004], who states that "*[...] in many real world applications, more complex networks of concepts are needed*". Therefore, the *internal* and *external* structure must be described. Relations among classes are interpreted as the subset of the product of the domain [Euzenat and Shvaiko, 2007]. Whether the modeler creates a separate class, or a qualified attribute relation, is just a matter of their modeling style.

[28] `http://protege.stanford.edu/doc/users_guide/` (last accessed February-28-2011).
[29] `http://www.daml.org/ontologies/` (last accessed February-17-2011).

7. *Define constraints:* The vocabulary (e.g., OWL DL) provides constructs to constrain the meaning of ontology entities (e.g., quantifier restrictions, cardinality restrictions, domain and range axioms). Such formal axioms provide a well-formed use of terms.

   The steps $1 - 7$ specify the *terminology* at the ontology layer. This terminology includes all essential concepts of the domain, the classification, the taxonomy, relations among concepts, and axioms to constrain interpretation.

8. *Create instances:* In the last design step the individual instances of the classes are created. When the modeler creates those instances, they think of them as the individuals in the domain of discourse. "*Individual instances are the most specific concepts represented in a knowledge base*" [Noy and McGuinness, 2001]. Therefore, they reside at the lowest level of granularity in the representation, at the information or data layer.

The problem is that there are several options to design a domain. In the OWL Web Ontology Language Guide[30] it is denoted that "*the development of an ontology should be firmly driven by the intended usage*". In a multitude of research works [Bontas-Simperl and Tempich, 2006, Grüninger and Fox, 1995, Janiesch, 2010, Noy and McGuinness, 2001, Park and Woo, 2007, Ramesh and Dhar, 1992, Smart and Engelbrecht, 2008, Uschold and Grüninger, 1996] the authors point out two rules in the context of ontology design: firstly, there exists no single correct ontology design methodology; and secondly, the potential usage of the domain ontology, as well as the engineers' understanding and personal view of the domain will undoubtedly affect ontology design decisions.

## Ontology Design Process

The ontology design process is principally *collaborative* and *iterative*. Existing knowledge bases are influenced by the work of many people of different disciplines. Gómez-Pérez et al. [2003] differentiate between *domain experts*, who provide the knowledge about the domain to be modeled, *ontology engineers*, who have experience in the fields of knowledge representation, ontology languages, and tools (e.g., Protégé, TopBraid Composer[31], Chimaera[32]), and *users*, who reuse ontologies for a certain purpose. In the phase of the design process engineers and experts design an abstract model (i.e., an ontology) of some phenomenon of the world. The area of interest, in which the concepts as well as the relations that hold among those exist, is known as the *domain of interest* or *domain of discourse*. The term *design* denotes the activities that lead to the development of an ontology. Figure 3.1 presents such a design process, documented in separate steps ①-⑦, in which multiple perspectives of a matter are condensed into a shared conceptualization. An ontology engineer and a system analyst in collaboration with domain experts build a domain ontology from scratch. They represent their view of the real-world domain using ontology entities (e.g., classes and relations) for its description. Steps ①-③: their view is based on the purpose or goal for that the ontology is modeled. Step ④: by the steps ①-③ the *context of purpose* or *domain context* is defined, in which ontology entities are used. Step ⑤: the ontology

[30] `http://www.w3.org/TR/owl-guide/` (last accessed August-1-2011).

[31] `http://www.topquadrant.com/products/TB_Suite.html` (last accessed July-6-2011).

[32] `http://www.ksl.stanford.edu/software/chimaera/` (last accessed February-13-2011).

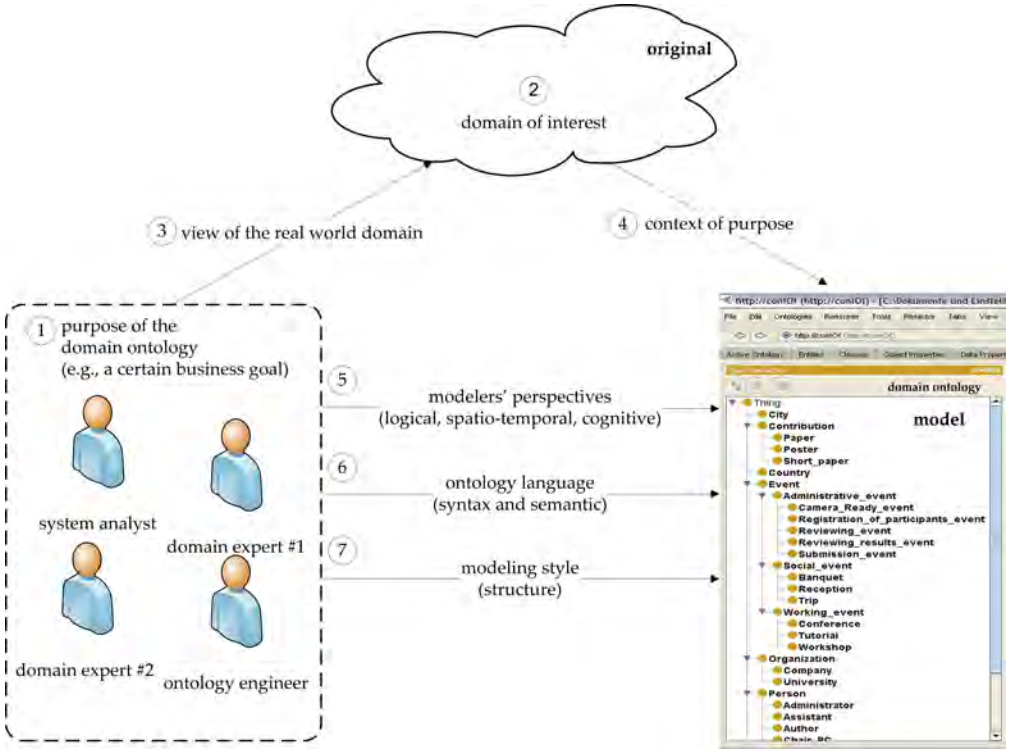



Figure 3.1: Ontology design process from scratch.

authors' view of the domain is specified by three perspectives: logical, spatio-temporal, and cognitive (Benerecetti et al. [2001], cf. Section 2.2). Step ⑥: when a group starts to conceptualize a certain domain it should agree on some shared representation forms, e.g., an expressive ontology language. The language provides the syntax and semantics to formalize domain-relevant knowledge. Step ⑦: finally, the engineers' modeling style or modeling conventions (Chalupsky [2000] and Klein [2001], cf. Section 1.2) influence the structure of the domain ontology.

Ontology design is not an absolute technical or logical process, rather it is a *socio-technical process* where a socio-technical context exists [Janiesch, 2010]. Experiences show that the bottleneck of building shareable ontologies mainly lies in the social process [Benjamins and Fensel, 1998]. There is a complex relationship among the ontology authors, the tools and the techniques in compliance with the artifacts associated with knowledge acquisition. The results of the study conducted by Smart and Engelbrecht [2008] (cf. Section 1.2) indicate that the major role in the design process is played by the modelers' internal representation of knowledge. A domain ontology is a memory map of reality; therefore the engineers "*should take into account the (entities) relation to the real world entities they are referencing, i.e., their meaning, as well as their purpose in the real world, i.e., their usage*" [Ehrig et al., 2004]. Additionally, to the entities' meaning in the real-world the specified domain context has an impact on their usage, as well as the engineers' experience in using language constructs. For instance, in the Encyclopedia

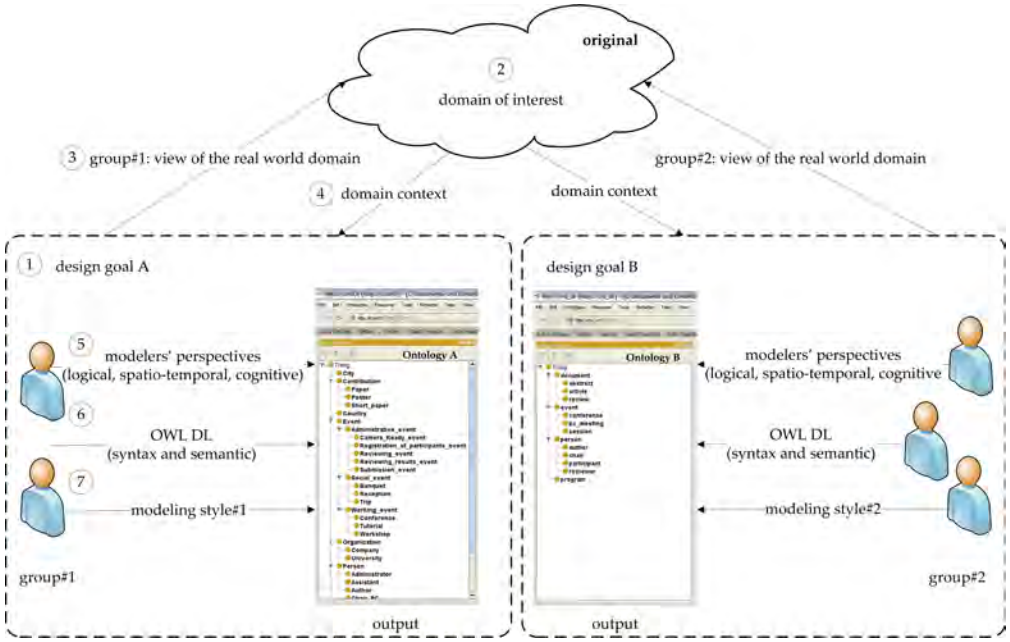



Figure 3.2: Example of two ontology design processes (group#1, group#2) with obviously different underlying design decisions resulting in dissimilar structures.

Britannica[33] the term "author" is defined as "*one who is the source of some form of intellectual or creative work; especially, one who composes a book, article, poem, play, or other literary work intended for publication*". There may be a difference in the meaning of this term depending on whether the ontology is to be modeled for the purpose of describing a software tool for conference organization support, or for describing a literature circle. In each of these scenarios "author" is used in different contexts with different information significance (i.e., relevance). In order to guide a decision process for determining the design goals, *competency questions* [Grüninger and Fox, 1995] are an efficient aid. They can help to define the requirements in the form of questions, which the ontology must be able to answer. Beside determining the scope, competency questions can guide the ontology authors to select the most important concepts and relations among them [Staab et al., 2001]. There exist *informal* and *formal* competency questions, which we introduce in detail by a concrete example in Section 6.2.

These are only a few reasons why ontology design is a rich knowledge process. It externalizes a valuable amount of domain-related background knowledge [Bontas, 2005]. Thus, this *process knowledge* should be made exploitable for users in ontology alignment. That would facilitate users in a better understanding of the sources, which would reduce time and cost expenditures when aligning them. The rationales underlying the design process are the engineers' *design decisions* [Gruber, 1995]. Such design decisions reflect a modeler's *mental state* at design time [Ramesh and Dhar, 1992]. The results of our evaluation survey, similarly to that of Smart and Engelbrecht [2008], indicate that beside specific design goals this mental state is ad-

[33] `http://www.britannica.com` (last accessed September-17-2011).

ditionally influenced by: previous experiences in ontology development, personal preferences, skills, socio-cultural impact, and the individual use of knowledge acquisition techniques. However, Park and Woo [2007] point to the lack that design decisions are typically not documented, and therefore this process knowledge is not on-hand for users, since it is implicitly encoded in the ontology's structure. The problems which may occur by the absence of such knowledge are illustrated by a graphical example. Figure 3.2 shows the design processes of two engineering groups, which both describe the same domain of interest with their own respective modeling focus on that domain and a particular design goal (design goal A vs. design goal B). Each goal forms the basis for determining the context in which the domain is to be modeled. The figure shows that obviously, even though the groups describe the same domain, they make different design decisions, because the ontologies have dissimilar taxonomic structures, which are also be impacted by different modeling styles. These differences may lead to structure- as well as pragmatic-based heterogeneity problems between Ontology A and Ontology B resulting in a mismatch when aligning these sources. However, there exist neither an "overall" context, nor a "global" perspective when different ontology authors describe the same domain of interest. Instead, referred to Saxe's theory of "variety of perspectives", there are different dimensions of context-dependent representations, according to Ehrig [2007], who states that "*there are many contexts in which an ontology can be considered, from the point of view of determining the similarity, the most important one is the application context*". The term "application context" is synonymous to the term "domain context" used in the graphic.

## 3.2 Ontologies in OWL

An ontology is expressed in a specific ontology language. Together, the vocabulary and the structure of an ontology provide a conceptual framework for analysis and information retrieval in a domain of interest [Gašević et al., 2006]. A variety of languages allow users to write explicit, formal conceptualizations of domain models. The W3C[34] has approved a standard vocabulary for representing ontologies,—OWL[35]. OWL is an ontology language for publishing and sharing ontologies on the Semantic Web. It provides formally defined meaning for its entities (e.g., classes, properties, axioms). There are three sub-languages of this vocabulary: (1) OWL Lite, (2) OWL DL, (3) OWL Full, each of which is more expressive than its predecessor. The usage of OWL ontologies supported by DL makes it feasible to encode anything which is conceptual in an expressive logical form. We focus on OWL DL throughout this thesis, because of its expressiveness compared to OWL Lite, its decidability compared to OWL Full, and due to the fact that DLs are a widely agreed standard for describing terminological knowledge. OWL DL places a number of constraints on the use of its constructs. The language is based on various features: classes and subsumptions, properties, type constraints, and others. When we further discuss ontologies, we mean domain ontologies expressed in OWL DL. Generally, an ontology in OWL is a set of axioms and facts by which the knowledge of a domain can be captured for a certain purpose [Euzenat and Shvaiko, 2007]. It describes the concepts as well as the local behavior of the relevant instances (i.e., individuals), which ontology authors are interested in.

[34] W3C = World Wide Web Consortium, `http://www.w3.org/` (last accessed February-28-2011).

[35] OWL = Web Ontology Language

"*An ontology together with a set of individual instances of classes constitutes a knowledge base*" [Noy and McGuinness, 2001].

Domain concepts can be described using the construct `owl:Class`. "*Classes are a concrete representation of concepts*" [Horridge et al., 2007]. OWL classes are interpreted as sets that contain individuals. Generally, classes are the main entities of an ontology [Euzenat and Shvaiko, 2007]. In OWL the class `owl:Thing` ($\top$) represents the set containing all individuals. Therefore, each class that is defined by an ontology engineer is a subclass of `owl:Thing`. The opposite of this root class is `owl:Nothing` ($\bot$), which is an empty class and defined as subclass of all classes. Classes can be organized in the form of a *concept hierarchy*, as super- and subclasses. Such a hierarchy is also known as *taxonomy*. Subclasses are specializations of their superclass, and inherit all features of that class. They are related by the transitive `rdfs:subClassOf` relation. For instance, if *workshop* is a subclass of *working event* and *working event* is a subclass of *event* then *workshop* is a subclass of *event*.

There are conditions (*necessary* & *sufficient*) that must be satisfied by an individual for class membership and class assignment. Both conditions are needed for reasoning (e.g., to automatically compute a class hierarchy). A class described by necessary conditions is denoted as *primitive* or *partial class*. Necessary conditions ($\sqsubseteq$) are needed to describe the membership of a class; "*if something is a member of this class then it is necessary to fulfill these conditions*" [Horridge et al., 2007]. That knowledge is not sufficient to determine if something (i.e., a random individual) fulfills the necessary conditions of a class then it is (automatically) a member of that class. To make this feasible *necessary conditions* have to be converted to *necessary* & *sufficient* ($\equiv$) conditions [Horridge et al., 2007]. A class defined by these conditions ($N\&S$) is denoted as *defined* or *complete class*. By a defined class it can be additionally said that if any individual satisfies the conditions of class membership then it must be a member of that class. Additionally, for separating a group of classes, they can be disjoint from one another. This ensures that an individual, which is asserted to one class of the group cannot be a member of any other class in that group.

There are other relations in OWL DL, in order to assert specific facts about individuals, or general facts about the class' members [W3C, 2004]. Such relations can be expressed by *properties*. There exist two main types: `owl:DatatypeProperty` and `owl:ObjectProperty`. The former relates individuals to data values (e.g., integer, string). Object properties are *binary relations* that assert a certain (labeled) type of relationship between individuals. They may have various characteristics for the purpose of reasoning, which are specified as: *functional*, *transitive*, *symmetric*, *inverse*, and *inverse functional* (cf. Section 3.3). OWL 2 [W3C, 2009] provides more characteristics by which the meaning of properties can be enriched even more. With regard to the method of our approach (cf. Section 6.1) we emphasize in particular that "*each object property may have a corresponding inverse property*" [Horridge et al., 2007]. For example: if the property *writes* has a corresponding inverse property *writtenBy*, and if *writes* link an author `a` to a certain contribution `b`;

```
<owl:ObjectProperty rdf:about="#writes">
    <rdfs:domain rdf:resource="#Author"/>
    <rdfs:range rdf:resource="#Contribution"/>
</owl:ObjectProperty>
```

then because of the inverse property we can infer that contribution `b` is written by author `a`;

```
<owl:ObjectProperty rdf:about="#writtenBy">
  <rdfs:domain rdf:resource="#Contribution"/>
  <rdfs:range rdf:resource="#Author"/>
  <owl:inverseOf rdf:resource="#writes"/>
</owl:ObjectProperty>
```

It is also possible to model hierarchies of properties, in the form of super- and subproperties. However, in OWL DL it is not possible to mix object and datatype properties; that means if a relation is defined as object property it cannot be a datatype property in the same ontology, and vice versa. An object property can have a *domain* (`rdfs:domain`) and a *range* (`rdfs:range`), which are class constraints to be checked. The property is then restricted by those axioms in such a way that it relates individuals of the domain to individuals of the range. Thus, these axioms constrain the meaning of the terms used in the vocabulary. There can also be multiple classes specified as the domain or range of an object property. Multiple domain classes mean that the `rdfs:domain` of the property is then the intersection of that classes; and similarly for range [W3C, 2004].

There may be other *restrictions* on properties. "*A restriction is a kind of class, in the same way that a named class is a kind of class*" [Horridge et al., 2007]. The restriction describes an *anonymous (unnamed) class*, which contains all the individuals that satisfy that restriction. There are three main categories [Horridge et al., 2007]:

1. *Quantifier restrictions*, specified by the *universal quantifier* ($\forall$) or *existential quantifier* ($\exists$), put constraints on relations that individuals participates in. The $\exists$-quantifier describes that *at least one* kind of relation along the specified property must exist from individuals of a class to individuals of a specific class, whereas the $\forall$-quantifier describes the set of individuals that *only* have relations along the restricted property to individuals of a specific class.
2. *Cardinality restrictions* specify the number of relationships an individual may participate in for a given property.
3. *HasValue restrictions* can be defined for a set of individuals that have *at least one* relationship along a specified property to a *specific individual*.

We described the constructs on the basis of the tutorial to build OWL ontologies using Protégé as ontology editor presented by Horridge et al. [2007] and the OWL Web Ontology Language Guide [W3C, 2004]. Further contributions of an ontology development methodology are presented by Gruber [1995], Noy and McGuinness [2001], Staab and Studer [2004], Uschold and Grüninger [1996], and Gašević et al. [2006].

## 3.3 Ontology Design Scenarios

The example ontologies, which we use in the following for the design scenarios, both describe the same domain of interest. In order to avoid unwanted direct or indirect influence on our part, we use example ontologies of the OAEI[36] 2009 evaluation campaign associated to the ISWC[37] ontology matching workshop. The domain of interest is a *software tool for conference organization support*. The first tool is the *Conference Management Tool*[38]. It is a Web-based event management system that was developed to support the organization of conferences, workshops, congresses, and seminars. This tool provides the following features: registration, administration, and invoicing of participants; the submission and review process of contributions; the scheduling of the conference program; and other organization tasks. The second tool is the *Conference Reviewing System*[39]. The purpose of this system is to make it feasible for users to manage all their conference reviewing and submission activities from one central location. The tool provides multiple role support to fulfill this requirement. For instance: chair for one conference, reviewer in another one, author in a third one. All these roles are manageable with a single account. Further features are: create review forms online, invite PC[40] members, manage submissions and reviewers, live Internet PC meetings, and end-to-end support for chairs.

The *Conference Management Tool* is described by the `confOf`[41] ontology alias $O_A$, and the *Conference Reviewing System* is described by the `crs_dr`[42] ontology alias $O_B$. The engineers of both ontologies use the same knowledge representation language (OWL DL) to describe the objects and relations among them. The example ontologies are quite suitable for our purposes due to their heterogeneous character of origin, since $O_A$ shows a deep and detailed concept hierarchy (three levels), whereas $O_B$ shows a shallow one with fewer classes at two levels. We will refer to these ontologies throughout the thesis whenever exemplarily needed. The snapshot, presented in Figure 3.3, makes visible the difference between the taxonomies of $O_A$ at the left side and $O_B$ at the right side. This form of heterogeneity causes structural mismatch when aligning these ontologies.

In the subsequent sections, we make a brief description of the two ontologies in natural language, and DL formalism. We interpret a few classes and their relations to other classes. Additionally, we explain certain object properties with their `rdfs:domain` and `rdfs:range` constraints. We select those classes and properties that are relevant for the application scenarios used in our evaluation survey (cf. Section 7.2). There are no property hierarchies within both ontologies. We omit restrictions on properties, *closure* and *covering* axioms, as well as instantiation.

[36] Ontology Alignment Evaluation Initiative, `http://oaei.ontologymatching.org/2009/conference/` (last accessed March-2-2011).

[37] $8^{th}$ International Semantic Web Conference, Fairfax (VA US), `http://iswc2009.semanticweb.org/` (last accessed March-2-2011).

[38] `http://www.conftool.net/en.html` (last accessed March-3-2011).

[39] `http://www.conferencereview.com/index.asp` (last accessed March-3-2011).

[40] PC = Program Committee

[41] `http://oaei.ontologymatching.org/2009/conference/data/confOf.owl` (last accessed June-12-2011).

[42] `http://oaei.ontologymatching.org/2009/conference/data/crs_dr.owl` (last accessed June-12-2011).

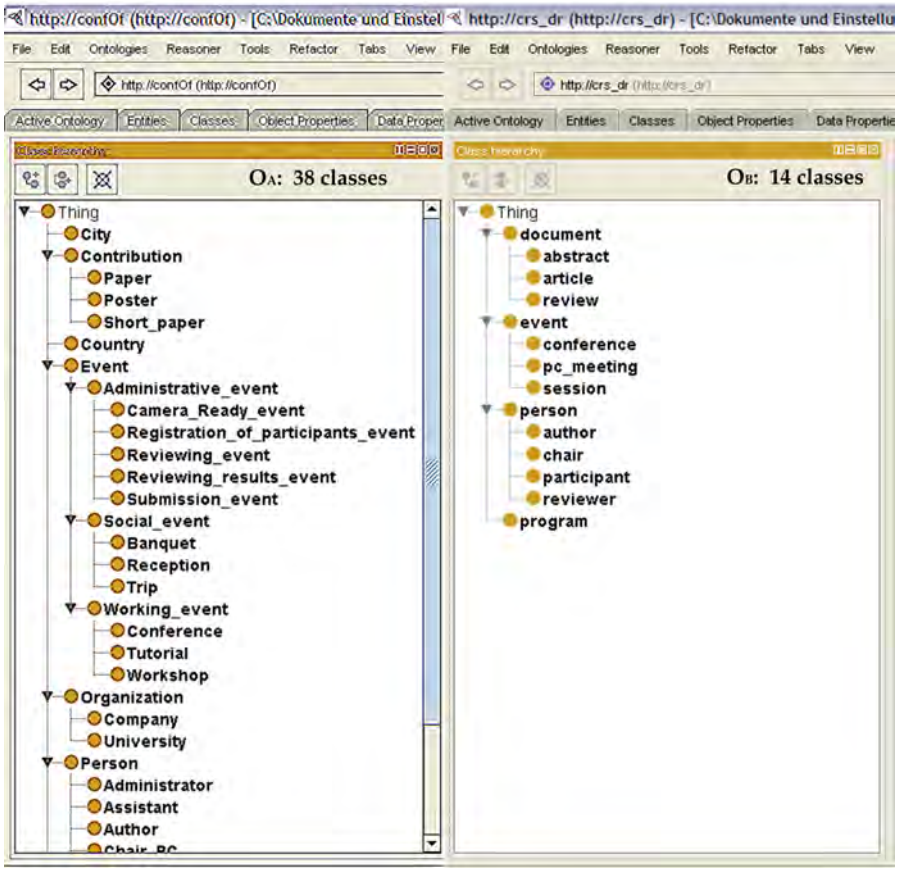



Figure 3.3: Composed snapshot of the Protégé editor where two differently structured ontologies (`confOf`, `crs_dr`) are presented that describe both the same domain of interest.

### Domain Ontology A: `confOf`

The `owl:Ontology` $O_A$ consists of a large number of classes (38), which are arranged in three hierarchy levels. The `owl:Class` *Person* contains the subclasses *Author*, *Administrator*, *Assistant*, *Chair_PC*, *Member_PC*, *Participant* (with a few subclasses more), *Scholar*, *Science_Worker*, and *Volunteer*. For instance, the `owl:Class` *Author* contains all the individuals that are authors in the domain of interest. The subclasses are siblings at the same level of generality, and not disjoint with each other. This means that individuals, which are members of *Author* can also be members of *Administrator*, or *Chair_PC*. Anyway, they are all `rdfs:subClassOf` *Person*. The `owl:Class` *Contribution* is disjoint of the `owl:Class` *Person*, which are both on the same hierarchy level. It has the subclasses: *Paper*, *Poster*, and *Short_paper*. These subclasses are disjoint. This means that if a certain contribution is asserted to the class *Paper* then it cannot be a member of the other classes in that group, too. For instance: authors write contributions, which can be papers, or posters, or short papers. If there is made an assertion of the form: `a` *writes* `b`; then it will be inferred that `a` is a member of the `owl:Class` *Author*, and that `b` is a member of the `owl:Class` *Contribution*. Inversely, a contribution is written by an author (`b` *writtenBy* `a`). The fact that a contribution deals with a topic is (formally) expressed by the logical statement (i.e., proposition):

```
<owl:ObjectProperty rdf:about="#dealsWith">
    <rdfs:domain rdf:resource="#Contribution"/>
    <rdfs:range rdf:resource="#Topic"/>
</owl:ObjectProperty>
```

The specified domain and range sets of the `owl:ObjectProperty` *dealsWith* are the `owl:Class`: *Contribution* and *Topic*. Additionally, the ontology authors had their focus on events when modeling the domain. There are: *Administrative_event*, *Social_event*, and *Working_event*. All of these are `rdfs:subClassOf` *Event*, and have further subclasses. Each working event has at least one ($\exists$) administrative event, and a certain topic. The engineers modeled that an administrative event follows another administrative event, as well as it can be parallel with an administrative event. Table 3.1 presents an excerpt of the aforementioned classes and their relations to other classes at the schema level (i.e., ontology layer). We use the DL formalism as presented by the *Knowledge Web Consortium* [Kno, 2007].

| **DL formalism** | **Interpretation** |
|---|---|
| $Contribution \sqsubseteq \top$ | the `owl:Class` ***Contribution*** is subclass of `owl:Thing` |
| $Paper \sqsubseteq Contribution$ | ***Paper*** is subclass of ***Contribution*** all individuals of ***Paper*** are individuals of ***Contribution*** without exception |
| $Poster \sqsubseteq Contribution$ | ***Poster*** is subclass of ***Contribution*** |
| $Short\_paper \sqsubseteq Contribution$ | ***Short_paper*** is subclass of ***Contribution*** |
| $Paper \sqsubseteq \neg Poster$ | if something is a member of the `owl:Class` ***Paper*** then this implies that it is not a member of the `owl:Class` ***Poster*** |
| $\top \sqsubseteq \forall writes^{-}.Author$ | the `owl:ObjectProperty` ***writes*** has the domain set of the `owl:Class` ***Author*** |
| $\top \sqsubseteq \forall writes.Contribution$ | the range set of ***writes*** is the `owl:Class` ***Contribution*** |
| $\top \sqsubseteq \forall writtenBy^{-}.Contribution$ | the domain set of the `owl:inverseOf` ***writtenBy*** property is the `owl:Class` ***Contribution*** |
| $\top \sqsubseteq \forall writtenBy.Author$ | the range set of the `owl:inverseOf` ***writtenBy*** property is the `owl:Class` ***Author*** |
| $\top \sqsubseteq \forall dealsWith^{-}.Contribution$ | the domain set of the `owl:ObjectProperty` ***dealsWith*** is the `owl:Class` ***Contribution*** |
| $\top \sqsubseteq \forall dealsWith.Topic$ | the range set of ***dealsWith*** is the `owl:Class` ***Topic*** |
| $\top \sqsubseteq \forall follows^{-}.Administrative\_event$ | the domain set of the `owl:ObjectProperty` ***follows*** is the `owl:Class` ***Administrative_event*** |
| $\top \sqsubseteq \forall follows.Administrative\_event$ | the range set of ***follows*** is the `owl:Class` ***Administrative_event*** |

Table 3.1: Domain ontology $O_A$ (*Conference Management Tool*)

### Domain Ontology B: `crs_dr`

The `owl:Ontology` $O_B$ has significantly fewer classes (14) with only two levels of hierarchy. There are four domain specific root classes: *document*, *event*, *person*, and *program*, which are disjoint with each other. The class *document* has the subclasses: *abstract*, *article*, and *review*, which are also disjoint. This means that nothing can be both, an abstract and an article. For instance: abstracts are contained in articles, but they are not a `rdfs:subClassOf` the `owl:Class` *article*. The authors of $O_B$ viewed an abstract not as a specialization of *article*. The `owl:ObjectProperty` *has_abstract* relates individuals of the `owl:Class` *article* to individuals of the `owl:Class` *abstract*. This implies that two individuals of these classes may be related by the *has_abstract* property. Additionally, the object property is `functional`, that means it is *single valued*. Inversely, an *abstract* is part of an *article*. The *part_of_article* property is `functional` and the inverse property of *has_article*. In the OWL Web Ontology Language Guide [W3C, 2004] a functional and an inverse functional property are formally defined as follows:

**Definition 1.**
$P\,(x,\ y) \wedge P\,(x,\ z) \Rightarrow y = z$, if a property is functional, for a given individual, there can be at most one individual that is related to the individual via the property [W3C, 2004].

**Definition 2.**
$P\,(y,\ x) \wedge P\,(z,\ x) \Rightarrow y = z$, if a property is inverse functional then it means that the inverse property is functional [W3C, 2004].

The individuals of the class *author* are as well members of the `owl:Class` *person*. This implies that *author* inherits all the features (e.g., name, address) of *person*. The engineers described an author as a person, who writes articles and submits them to conferences. An article can be written by more than one author. The `owl:Class` *person* has further subclasses: *chair*, *participant*, and *reviewer*, all of which are pairwise disjoint. This means that an author, who submits an article to a certain conference cannot be at that conference a chair, or a reviewer, too. This is the engineers' subjective view of the domain, which has no claim of general validity, since there are conferences where an author, or co-author of a contribution has also the role of a reviewer of other contributions.

| **DL formalism** | **Interpretation** |
|---|---|
| $abstract \sqsubseteq document$ | *abstract* is subclass of *document* |
| $article \sqsubseteq document$ | *article* is subclass of *document* |
| $review \sqsubseteq document$ | *review* is subclass of *document* |
| $document \sqsubseteq abstract \sqcup article \sqcup review$ | to be a *document*<br>it is necessary to be<br>an *abstract*, or an *article*, or a *review* |
| $author \sqsubseteq person$ | *author* is subclass of *person* |
| $\top \sqsubseteq \forall has\_abstract^{-}.article$ | the domain set of the<br>`owl:FunctionalProperty` *has_abstract*<br>is the `owl:Class` *article* |
| $\top \sqsubseteq \forall has\_abstract.abstract$ | the range set of the<br>`owl:FunctionalProperty` *has_abstract*<br>is the `owl:Class` *abstract* |
| $\top \sqsubseteq \forall part\_of\_article^{-}.abstract$ | the domain set of the<br>`owl:InverseFunctionalProperty`<br>*part_of_article* is the `owl:Class` *abstract* |
| $\top \sqsubseteq \forall part\_of\_article.article$ | the range set of the<br>`owl:InverseFunctionalProperty`<br>*part_of_article* is the `owl:Class` *article* |
| $\top \sqsubseteq \forall writes\_article^{-}.author$ | the domain set of the<br>`owl:ObjectProperty` *writes_article*<br>is the `owl:Class` *author* |
| $\top \sqsubseteq \forall writes\_article.article$ | the range set of the<br>`owl:ObjectProperty` *writes_article*<br>is the `owl:Class` *article* |
| $\top \sqsubseteq \forall article\_written\_by^{-}.article$ | the domain set of the<br>`owl:inverseOf` *article_written_by* property<br>is the `owl:Class` *article* |
| $\top \sqsubseteq \forall article\_written\_by.author$ | the range set of the<br>`owl:inverseOf` *article_written_by* property<br>is the `owl:Class` *author* |

Table 3.2: Domain ontology $O_B$ (*Conference Reviewing System*)

## 3.4 Concluding Remarks

In this chapter we discussed the phases of ontology development, and we described an ontology design process from a socio-technical perspective. We gave a brief overview to that OWL-constructs, which are necessary for our approach, its implementation, and which are used in the example ontologies (`confOf` ($O_A$) and `crs_dr` ($O_B$)) that we take for explanation. The graphical comparison of these ontologies, presented in Figure 3.3, shows their scope and granularity. The ontologies are heterogeneous in their structures, since ontology $O_A$ has a deep and detailed concept hierarchy with many `rdfs:subClassOf` relations, whereas ontology $O_B$ deals with a wide range of classes, and therefore a lot of semantic connections (`owl:ObjectProperties`) among that classes. We discussed that a design process is based on the engineers' design decisions, which are reflected in *usage patterns* (e.g., how they relate entities in order to describe the domain for a certain purpose). Such patterns are inherent (i.e., implicitly encoded) in the ontology's structure; and therefore not immediately visible to users in ontology alignment. Additionally, design decisions are affected by the engineers' experience in modeling ontologies, their personal preferences, skills, as well as their background knowledge of the domain. Different design decisions may lead to several forms of heterogeneity resulting in a mismatch between two ontologies; since "*different model conceptualizations are difficult, if not impossible problems to solve in ontology reuse*" [Smart and Engelbrecht, 2008].

Ontologies are not limited to conservative definitions, which are definitions in the traditional logic-based sense that only introduce terminology and do not add additional knowledge about a domain [Enderton, 1972]. Users need such *informal* knowledge to estimate the quality of ontologies when aligning them. Often, they are not familiar with the ontology to be evaluated, or "*the ontology is utterly too complex to be read through by humans*" [Bontas, 2005]. Domain-related background knowledge about the original purpose of an ontology and its design process can give exploitable hints about its appropriateness for the alignment with other ontologies. Such hints would be useful for assisting the users' decision-making process; for instance, by supporting them with evidence of the engineers' implicitly made assumptions about domain concepts (e.g., their relevance in a certain context). Therefore, we present in the next chapter the idea to implement an evidence-based unidirectional communication model from the original developers to users.

CHAPTER 4

# Cognitive Design Methodology

The drawbacks posed in the previous chapters highlighted the limitations when aligning ontologies. That makes it necessary for us to follow a better integrated user support in ontology alignment that already starts when developing ontologies. Such an early consideration of "alignment support" is crucial in our approach. For this purpose we introduce *CoMetO*,—a *cognitive design methodology for enhancing the alignment potential of ontologies*. In this methodology we consider context from a cognitive point of view, as discussed in the cognitive theories by Bouquet and Serafini [2000] (cf. Section 2.3) in order to obtain a broader, intentional view of the domain. We introduce a formalism,—the *modeling focus (MF)* by which the engineer's cognitive perspective related to certain contexts can be represented. Based on this concept we present the idea to implement an *evidence-based (unidirectional) communication model* from engineers to users, which is mainly influenced by the *relevance-based inferential model* of verbal communication.

The main objectives of our approach are: to fulfill a higher level form of meaning negotiation as proposed by Bouquet et al. [2002], and pointed out by Noy and Musen [2002]; to reflect the modeler's decisions made at design time as discussed by Ramesh and Dhar [1992], Bontas [2005], Park and Woo [2007]; and to implement a method by which semantic content associated to context can be related to the ontology layer, as proposed by Janiesch [2010].

## 4.1 Terminology

Often, no agreement is found on the exact meaning of terms. The listed terms and their definitions are consolidated as they are used, according to their specified meaning, in this thesis. We try to be consistent as far as possible with definitions in other publications. Terms that are sufficiently defined in one of the previous sections are not repeated here.

**Concept and class:**

When we discuss domains or ontologies we generally use the term concept. When we focus on a domain ontology, which describes a certain domain of interest, we use the term class. We make

use of OWL DL as ontology language. Therefore, we mean an `owl:Class` when we utilize the term class.

**Contexts:**

We introduce two contexts involved in the socio-technical context of a development task, which we specify as:

a. **Domain context:** We define as *domain context* ($C_D$) the requirements of the specific purpose to fulfill its design goals, linked to the domain for that the ontology is modeled. So, we meet the demands proposed by Gruber [1995] and Grüninger and Fox [1995] (cf. Section 3.1) by considering such a context. We use the term "domain context": firstly, because the focus of this thesis is on domain ontologies; secondly, to avoid overlaps among the different kinds of ontologies (e.g., domain ontology, application ontology, task ontology, etc.) and their specific tasks; and finally, because the term "application context" is often used related to the users motivation of reusing ontologies, as described by Bontas [2005] (cf. Section 2.2). The term "domain context" is used akin to the term "application context" discussed by Ehrig [2007] and Janiesch [2010].

b. **Modeling context:** Ontology engineers rely on their experience, skills, preferences, education, etc. when modeling a domain of interest (cf. Section 1.2). An ontology is structured either deep, or more shallow in a way (cf. Figure 3.3 in Section 3.3). On the one hand, the chosen representation is a matter of the modeler's taste, which is mainly based on their experience in ontology engineering; and on the other hand, it is caused by the importance of a class, as introduced by Noy and McGuinness [2001]. We define the context to which we relate those modeling conventions as the *modeling context* ($C_M$). $C_M$ is mainly responsible for the structure of a domain ontology.

**Context-sensitive meaning or cognitive semantics:**

Elman [1989] defines that "*sensitivity to context is precisely the mechanism which underlies the ability to abstract and generalize*". The term combinations *context-sensitive meaning*, or *cognitive semantics* define the consideration of a logical statement's meaning in a certain context (i.e., its contextual effect). We need to distinguish between the intended meaning of a vocabulary (its semantics) and the intended meaning of the modeler, who uses the constructs of that vocabulary to describe domain concepts in order to fulfill a certain design goal. The latter is known as cognitive semantics, which conveys the entities' *meaning in use*.

**Logical statement:**

Propositions or sentences, which are modeled at the schema level of an ontology (ontology layer), are often denoted as *logical statements* in the literature. This term has to be distinguished from statements made at the information layer of an ontology, which contains instance data and no schema level information. A logical statement, when formulated as a sentence, is a well-formed formula with no free variables. It can be expressed by using existential ($\exists$) or universal ($\forall$) quantifiers, which are *truth-bearers*. The notion "truth-bearer" means that the statement in

the form of a declarative sentence is either *true* or *false* [Hitzler et al., 2008]. A proposition is well-formed consisting of atomic formulas, but without quantifier variables. In this thesis, when we use the term "(logical) statement" we mean a proposition.

**Modeler and user:**

There are many actors involved in the development and alignment of ontologies. We distinguish two major roles: *modeler* and *user*. The modeler's role is taken by the ontology authors (e.g., domain experts, ontology engineers), who are concerned with knowledge acquisition and the development task of ontology engineering. It is also not uncommon that one person has the combined role of both domain expert and ontology engineer. The user's role is taken by the end-users, who want to reuse (e.g., align) ontologies.

**Modeling focus:**

There exists a cognitive perspective on the domain as introduced by Benerecetti et al. [2001], which is akin to the mental state presented by Bouquet et al. [2002] in their cognitive theories of context, and the mental state discussed by Ghidini and Giunchiglia [2001] (cf. Sections 2.2, 2.3). The *modeling focus (MF)* is a representation formalism for such a perspective. It helps to join an engineer's logical and cognitive perspective on the domain when describing its concepts at the schema level. By the modeling focus the engineer gives information of the entities' meaning in use in certain contexts ($C_D$, $C_M$) to users. Its implementation facilitates a *use-conditional* form of meaning consideration as well as interpretation additionally to a *truth-conditional* one, as made practicable by first-order logic.

## 4.2 Background of CoMetO

Missing expert knowledge of the entities' usage in context may lead to problems (e.g., pragmatic heterogeneity) in the interpretation task caused by the lack of a user's capability to comprehend design decisions. This fact reveals that "*the intended use of entities has a great impact on their interpretation, therefore, matching entities which are not meant to be used in the same context is often error-prone*" [Euzenat and Shvaiko, 2007]. We assume that meaning interpretation involves more than merely identifying the semantics of assumptions explicitly expressed (e.g., by model theory-based techniques), since there exists an intended meaning depending on context. This context-based meaning is intended by the modeler, who uses vocabulary constructs (e.g., `owl:Class`, `owl:ObjectProperty`) to describe domain concepts, and is therefore not to be confused with the intended meaning of the vocabulary. The interpretation of such "meaning in use" makes additional (expert) background knowledge of the domain ontology viz. its design process necessary. Established alignment methods have the lack of deducing such intentional knowledge only, if ever, *ex post* without contextual evidence. Therefore, the burden of meaning interpretation is still on users [Hughes and Ashpole, 2004].

From our literature-based analysis we can state that unresolved problems, which may occur in the alignment process are: firstly, caused by the lack that such specific knowledge about the usage of entities is implicitly contained in the structure of ontologies, and therefore not

visible to users in the alignment task; and secondly, the scope of a user's interpretation is individual, as well as context-dependent resulting from their own (cognitive) perspective when aligning ontologies. With the latter problem we want to point out that the users' assumptions about a domain affect their meaning interpretation, because "*interpretation is an essential part what people know*" [Bouquet et al., 2002]. Thus, their interpretation may diverge with what the modeler intends it to be when they express statements, since there exist different world views. Currently, the interpretation task is only supported from a logic-based point of view by exploiting model-theoretic semantics by using set theory. Model-theoretic semantics provides the rules for interpreting the syntax of a language, which does not directly provide meaning. It only constrains the possible interpretation of what is declared [Euzenat and Shvaiko, 2007]. The cognitive perspective—which is necessary to exploit cognitive semantics is known, but not used for meaning interpretation. There exists no access to that knowledge even if users require it, as demanded by a participant in the course of the online survey conducted by Falconer et al. [2007] (cf. Section 1.2): "*it would be a great benefit to get into the brains of the original developers*". For this reason, we implement a procedural method in the first part of CoMetO by which modelers are supported to provide evidence of their cognitive perspective to users.

We keep the knowledge, contained in each ontology, *local*. Thus, we focus on domain ontologies autonomously when aligning them, similar to the approach introduced by Bouquet et al. [2002]. With such a "local view" of the sources we want to retain their particular model-based and cognitive semantics in order to avoid information loss. Therefore, we neither attempt to create a global (meta) ontology as presented by Bontas [2005], Bouquet et al. [2003a], nor, we implement an overall architecture where local ontologies are hierarchically organized as introduced by Giunchiglia [1992] (cf. Section 2.2). Also, we do not profit from a common grounding by upper level ontologies as SUMO[43] or DOLCE[44], which help in handling the disambiguation of multiple possible meanings of terms. There are alignment tools, which integrate such top level ontologies as cognitive aids. The idea is that it is easier to find correspondences between ontology concepts if the two sources extend the same reference ontology in a consistent way. Bouquet et al. [2002] argue that external interpretation schemas, which should help to define meaning, may lead to a loss of the engineers' innovative perspectives. However, in CoMetO neither a sense-based memory source, nor an upper level ontology are used in order to make cognitive semantics exploitable for meaning interpretation. In our approach the cognitive support of users is provided by the original modelers themselves.

"*When several users communicate the understanding cannot be ensured by the semantic embedding only*" [Euzenat, 2000]. Thus, we adapt a theory of *pragmatics* in order to enrich the ontology's (relational) structure with context-based *cognitive semantics* to provide—in combination with model-based *(logical) semantics*—a "complete package" for meaning interpretation as input in the alignment process. We hold that nobody can annotate such additional knowledge better than the ontology authors themselves. We act on the assumption that modelers are willing to share that knowledge. This assumption is based on: (1) we keep the modelers' additional

[43]Suggested Upper Merged Ontology, `http://www.ontologyportal.org/` (last accessed June-28-2011).

[44]Descriptive Ontology for Linguistic and Cognitive Engineering, `http://www.loa-cnr.it/DOLCE.html` (last accessed June-28-2011).

burden to express their cognitive perspective as low as possible in that we implement a method where they can annotate this knowledge to schema level entities by a simple point-and-click interaction in an ontology editor; (2) that annotation procedure is integrated in the process when creating a logical statement, which means that switching to another input mask or an extra tool is not required; and (3) modelers also benefit from sharing that knowledge, since they are in the role of a user when reusing existing ontologies (e.g., FOAF[45]) as starting point.

In CoMetO we consider the ontology's structure and omit instance data. We focus on the more general context of logical statements, rather than on situational details. We agree with Janiesch [2010] that the use of *situational context* is too detailed to facilitate a meaningful reuse. It fails to provide useful assistance for more than one situation associated with a certain individual. The degree of flexibility would be to high. Already, Giunchiglia [1992] states: "*contexts are not situations*". Therefore, we leave the instance level at the data layer and lift our approach one level up to the schema (i.e., model) at the ontology layer. We introduce two contexts in CoMetO: the domain ($C_D$) and the modeling context ($C_M$), which are interacting. This corresponds to the comment made by Giunchiglia [1992] that "*a knowledge base contains in general a set of interacting contexts*". The requirements for design decisions of the entities' usage are defined related to $C_D$, whereas the modeling style, which impacts the ontology's structure is defined related to $C_M$. The consideration of these two categories of context, involved in the task of ontology development, addresses an open issue pointed out by Janiesch [2010]. He describes the lack of general categories for structuring context dependency based on an analysis of the actual use of context in information and knowledge systems. He proposes to develop categories of context.

## 4.3 Approach of CoMetO

We view ontology development as a socio-technical design process, as introduced in Section 3.1. The ontology's purpose represents the goals to be satisfied by this process that in turn requires design decisions to satisfy these goals [Ramesh and Dhar, 1992]. If the purpose differs (e.g., due to certain business goals) the design goals and corresponding to them the design decisions will differ, too (cf. Section 3.3). Ramesh and Dhar [1992] define the designers' process knowledge as the linkage of design rationales to design artifacts. This knowledge captures *formal* and *informal semantics*. In CoMetO we consider it as the focused knowledge at design time, which involves the three perspectives introduced by Benerecetti et al. [2001] (cf. Section 2.2). We assume that *formal decisions* are related to the modeler's *logical perspective* in compliance with an ontology language, whereas *informal decisions* are related to their *cognitive perspective* on the domain. The logical or classical view posits the kind of knowledge modelers have when they describe domain concepts (e.g., person, author) well-formed by using the syntax and semantics of an ontology language. For instance, that is knowledge of some necessary and sufficient defining conditions for category membership (e.g., anybody who is related to the class *author* is a type of *person*). This means that the *semantic competence* refers to the modeler's intention to define well-formed axioms. Our experience leads us to the assumption, and the pilot study conducted by Smart and Engelbrecht [2008] (cf. Section 1.2) and a multitude of other works [Falconer

[45]FOAF = Friend of a Friend, `http://xmlns.com/foaf/spec/` (last accessed September-21-2011).

and Storey, 2007, Falconer et al., 2007, Noy and Musen, 2002] confirm it that design decisions are only to a small extent logical and deductive driven. In the design process modelers have a comprehensive access to contextual information: intuition, belief, goal, experience, and others, which are *cognitive abilities* [Norman and Draper, 1986] and as such *pragmatic features*. We assume when making the modeler's cognitive perspective on domain concepts visible to users that they would get a "bigger picture" of the sources as currently feasible.

In the approach presented by Ghidini and Giunchiglia [2001] (cf. Section 2.2) contexts are *mental images*. Bouquet and Serafini [2000] refer in their cognitive theories (cf. Section 2.3) to contexts as *cognitive constructs*, as such they are part of the modeler's mental state. Such a state is reflected in the modeler's behavior when developing a domain ontology. In CoMetO we consider two mental images: $C_D$ and $C_M$ (cf. Section 4.1). For instance, competency questions are an aid for determining the requirements to define $C_D$. They depend on the purpose of modeling the domain of interest. *Informal competency questions* determine the scope. "*They place the demands on an underlying ontology*" [Grüninger and Fox, 1995], but they are not expressed in a formal language. They are an informal justification of that domain to be modeled which provides an initial evaluation of ontology entities. For instance, the expressiveness of ontological commitments [Kalfoglou, 2000] can be evaluated by informal competency questions. *Formal competency questions* specify the informal questions. They help to formulate definitions to constrain the interpretation of entities. The statements which are satisfied by formal competency questions are included in the ontology's terminology in oder to make reasoning feasible [Grüninger and Fox, 1995].

The outlined works in Section 2.2 consider context from a *model-theoretic point of view*. Seen from this logical perspective a representation (e.g., domain ontology) is a set of assumptions about a domain explicitly expressed in a language ($\mathcal{L}$), which has usually the form of first-order logical theory [Guarino, 1998], more formally expressed in Euzenat [2001] and Euzenat and Shvaiko [2007]:

**Definition 3.**
$\mathcal{L}$ (ontology language) is a set of language constructs; $r$ (representation) is a set of expressions made using the constructs of $\mathcal{L}$; $\mathcal{I}$ is an interpretation function that maps each entity (e.g., class, relation) to a set in $\mathcal{D}$, which is called the domain of interpretation; the set of interpretations, which satisfy the assertions made in $r$ are models of $r$; $\mathcal{M}$ is the set of that models; $\delta_i$ ($i = 1, \ldots, n$) are axioms in $r$, which are satisfied by $\mathcal{I}$ if they meet certain conditions (e.g., that $\mathcal{I}(\delta_i)$ belongs to a subset of $\mathcal{D}$).

*Semantic quality* expresses the degree of correspondence between $r$ (the set of expressions) and the concepts of the domain to be modeled [Poels et al., 2005]. Such a representation is a logical form or well-formed formula and as that a structured set of concepts (conceptual representation). An assumption expressed in OWL in the form of a statement is a certain item of information about a domain. For example: we want to express that papers are contributions. We define two classes by declaring them as named classes by the labels: *Paper* and *Contribution*;

```
'Class(Paper)' ∈ r
'Class(Contribution)' ∈ r
```

in order to express that each paper is a contribution, we use the language construct `rdfs:subClassOf`.

$$\texttt{'SubClassOf(Paper Contribution)'} \in r$$

Generally, from a logic-based perspective the properties of a representation are logical properties, and semantics is the meaning of each axiom $\delta_j$ in $r \sqsubseteq \mathcal{L}$. The semantic representation of a statement deals with a sort of common meaning shared by every sentence of it. Bouquet and Serafini [2000] comment that semantics is defined only with respect to a single space, which looks like an objective representation of the world. Naturally, models are marked by their authors' focus on the domain, and therefore cannot claim to represent "purely" objective reality [Janiesch, 2010]. In an analogous way Sperber and Wilson [1995] comment that even if the members of a community use the same language, and converge to the same inferential abilities, the same is not true for their assumptions of the world. For instance, in order to specify the correspondence between the two classes $O_A$*:Contribution* and $O_B$*:Paper* an algorithm could use an external resource as WordNet, even though the intended meaning of these classes differs (since there is no evidence of their usage in context).

A well-formed sentence concerns the structure of a language $\mathcal{L}$, and therefore it has a fixed truth value with respect to that structure. Thus, formal logical operations are determined by such a structure. Maedche and Staab [2002] point to a drawback by using only the formal structure when aligning ontologies; they state that "*all real-world ontologies that we know of do not only specify its conceptualization by logical structures, but to a large extent also by reference to terms that are grounded through human natural language use*". Chomsky [1980] views knowledge of the language's use to achieve a certain purpose as *pragmatic competence*. From a logical perspective meaning is a function from assertions into truth values computed by deductive rules. In this context Carston [1998] criticizes that model-theoretic semantics "*as truth-conditional content is the minimal proposition expressed*".

Sampson [2007] defines pragmatic quality as "*the degree of correspondence between the ontology and the audience interpretation (the degree to which the ontology has been understood)*". We assume that for improving such a quality non-logical (informal) properties have to be expressed to interpret $r$ related to context. The differentiation of meaning interpretation can be described as follows: from a logical perspective the meaning of a concept is a function characterizing a set of objects *relative to context*, whereas from a cognitive perspective the meaning of a concept is intended by the modeler to characterize objects *in context*. Therefore, the latter is not exploitable by model theory, because truth conditions of a sentence are inapplicable when having a pragmatic-based (cognitive) view [Travis, 1997].

Decisions made during the design process lead to constraints on design objects (e.g., ontology entities) [Ramesh and Dhar, 1992]. Domain knowledge can be formalized by axioms. They specify the definitions of terms and constrain their interpretation. The process to define such *logical entries* can be guided by formal competency questions by which their completeness can be evaluated. In OWL DL axioms are a syntactic category beside entities (e.g., classes, properties, individuals) and expressions, which represent complex notions [Horridge et al., 2007]. They can be used to well-form statements at the schema level. Such a statement in its simplest form can be expressed by an `owl:ObjectProperty` with its certain domain/range sets (i.e.,

`ObjectPropertyDomain` and `ObjectPropertyRange`) (cf. Section 3.2). For instance: we define a simple example statement in a formal knowledge representation. We want to express the assumption that authors write contributions in the domain of interest. We need a binary relation *writes* and describe it by the domain *Author* and the range *Contribution*;

```
<owl:ObjectProperty rdf:about="#writes">
    <rdfs:domain rdf:resource="#Author"/>
    <rdfs:range rdf:resource="#Contribution"/>
</owl:ObjectProperty>
```

Visser and Cui [1998] comment that the meaning of a term (e.g., *Author*) within a logical statement depends on the meaning assigned to its primitives (e.g., `rdfs:domain`). The meaning of the example statement based on the vocabulary's semantic can be interpreted as follows: the object property *writes* has the domain set *Author* and the range set *Contribution*. This implies that if two individuals (`a`, `b`) are related by the binary relation *writes*: $R_{writes}$(`a`, `b`); then `a` is an instance of the class *Author* and `b` is an instance of the class *Contribution*.

The logical perspective leads to *semantic-based constraints* (e.g., formal specifications as in first-order logic) on expressed assumptions (e.g., statements). The cognitive perspective leads to *pragmatic-based (cognitive) constraints* on the usage of a statement related to the context in which it is processed. These constraints are *informal* and as such they have no effect on the logical aspects of the ontology. For instance, they can give information about the importance of a statement (i.e., its contextual effect) in the domain context ($C_D$). Generally, concepts are not equally important in the purpose-specific description of the domain. Therefore, they can be classified in more or less relevant to the goals, which have to be fulfilled to satisfy the ontology's purpose. Currently, established ontology languages neither provide certain constructs to define cognitive constraints on entities, nor they provide mechanisms to populate the metamodel with pragmatic-specific instances (e.g., characteristic factors as numerical values). Such context-based metalevel information does not talk about the domain, but describes domain information itself. For instance, that meta-information can be used to filter or rank concepts related to domain-specific information (e.g., to gather the core concepts). OWL DL provides annotations properties, but this construct is not expressive enough for our purpose.

Figure 4.1 illustrates the layers as introduced in the layer model (cf. Figure 2.1 in Section 2.3) and our idea to relate the cognitive perspective to a certain theory of pragmatics,—the *relevance theory* [Sperber and Wilson, 1995]. We adapt this theory for a cognitive-based improvement of ontology development and alignment. The consideration of meaning in a *use-conditional*, instead a *truth-conditional* form is facilitated by this theory. The distinction is that from a model-theoretic point of view (logical perspective) semantics is considered as truth-conditional by (formal) conceptual encoding, whereas from a pragmatic-based focus (cognitive perspective) semantics is considered as use-conditional by (informal) procedural encoding. The relationship between syntax and interpretations is defined by model-based semantics. Therefore, conceptual encoding makes automated consistency checking feasible. Pragmatics is dealing with the *effects of context*. For instance: there are two ontologies describing the same domain of interest using OWL DL. We assume that the purpose of one ontology is to arrange organizational procedures (e.g., workshops, events), whereas that of the other is to deal with submitted papers. The propo-

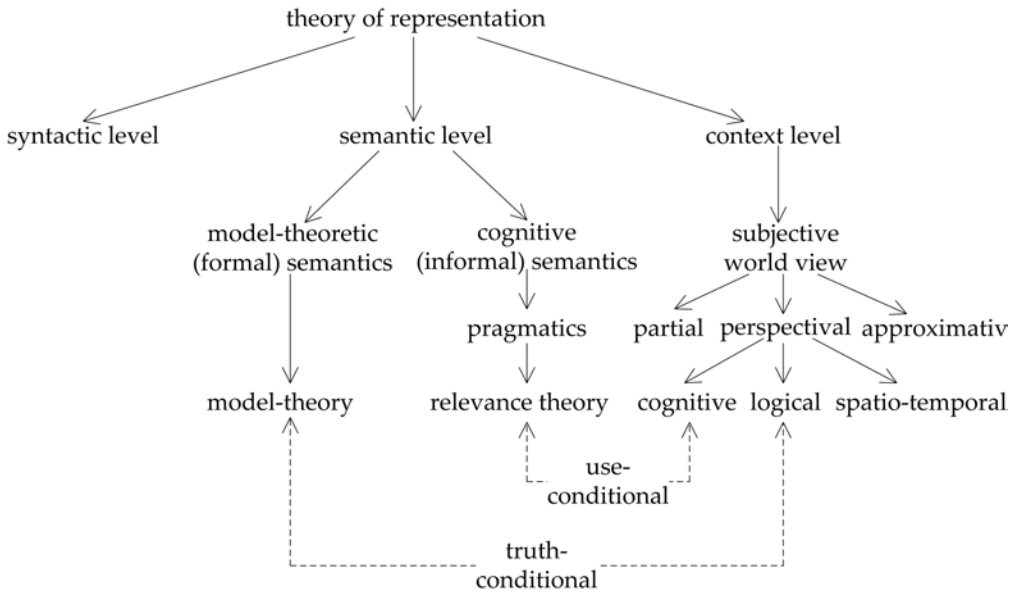



Figure 4.1: Extension of the layer model.

sition *writes* $\rightarrow$ (*Author*, *Contribution*), which is an expressed assumption in both ontologies, may differ in its intended usage insofar as it does not have the same *relevance in context*, and that may cause pragmatic heterogeneity problems as discussed in Section 1.2. The pragmatic-based condition "having some relevance in context" means that a statement with any contextual effect is relevant or important to some degree. That makes use-conditional inference to a *relevance-driven* processing, where *relevance* is a (mental) property of the inputs of cognitive processes [Sperber and Wilson, 1995]. Adapted to our approach this means that the modeler annotates (cognitive process) logical statements (the inputs) with additional meta-information about their relevance in a certain context ($C_D$), which can be inferred by the user through indicators.

We understand as an effective meaning negotiation a way to support users with all kinds of information available to get what the modelers have intended to convey for them. We assume that our adaptation of pragmatics viz. its relevance theory for the ontology's design process helps users in ontology alignment to analyze how entities are used to convey information related to context. Pragmatics is treated as "*code-like mental device*" [Carston, 1998] by which—as we assume—the semantic structure of an ontology can be supplemented. It considers both, code and context.

## 4.4 Pragmatics

*Pragmatics* is the study of linguistic or speech acts and the context in which they are processed. *Context* is the all-pervasive concept of pragmatics [Lab, 2006]. Generally, pragmatists distinguish among *sentences* and *utterances* or *propositions* in verbal communication. They view language as a code or grammar, "*which pairs phonetic and semantic representations of sen-*

*tences*" [Carston, 1998]. Pragmatists relate the interpretation of a sentence to grammar, whereas that of an utterance to pragmatics, because "*the semantic representation of sentences cannot be regarded as corresponding very closely to thoughts*" [Sperber and Wilson, 1995]. There are two major theories of communication in pragmatics: the *code* and the *inferential model*. In the first theory the focus is placed on a single model where communication is achieved by *encoding* and *decoding* messages. The code model is referable to *classic pragmatics*, which is mainly entrenched in the Western scholarly tradition [Lab, 2006]. In the second theory the communicator provides evidence of their intentions and the audience infers those intentions from that evidence. The evidence is a contextual information providing a precise purpose to the hearer. In the inferential model communication is achieved by *producing* and *interpreting* evidence; its domain is the speaker's informative communicative intention. We denote the representatives of the code model theory as *speech act theorists*, who take a classical view of pragmatics; and those of the inferential model theory as *relevance theorists*, who focus on pragmatics interdisciplinary. When we generally speak about the representatives of pragmatics we use the term pragmatists. Moreover, we use communicator/speaker, as well as audience/hearer as equivalent terms. Our aim is not to discuss pragmatics and its relevance theory in depth, but rather to present those parts of the subject that are necessary for understanding the approach made in CoMetO.

In our methodology we take a *relevance-theoretic* view of pragmatics, therefore we focus on the inferential and not the code model theory. In the inferential model semantics and pragmatics are distinct by *decoding* and *inference* [Carston, 1998]. "*Hearers are interested in the meaning of the sentence uttered only insofar as it provides evidence about what the speaker means*" [Sperber and Wilson, 1995]. It can be said the inferential model "*appeals to common sense*" [Potts, 2010]. The decoding process can be performed by autonomous linguistic systems (e.g., parser), whereas for the access to the implicit encoded knowledge a pragmatic inferential mechanism is necessary, which is *relevance-driven*. Pragmatic inference facilitates access to the contextual information that is intended by the speaker to be expressed in the proposition. This cognitive process is constrained by the *principles of relevance* or *informativeness* [Grice, 1975].

A common linguistic structure can be described by the generative grammar of the language akin to the semantic representation of ontologies (cf. Definition 3). Grammar takes account for purely linguistic properties, therefore non-linguistic properties are not considered (e.g., the speaker's intentions). An utterance contains both, *linguistic* and *non-linguistic* properties. The term utterance or proposition in verbal communication is a realization of the phonetic representation of a sentence. There is an interaction between linguistic structure and non-linguistic information when interpreting an utterance. Linguistic just like semantics is exploitable by grammar (i.e., code). Relevance theorists, in contrast to speech act theorists, argue that treating linguistic communication as the model of communication is a limited access to information. "*Different utterances of the same sentence may differ in their interpretation; and indeed they usually do*" [Sperber and Wilson, 1995]. Relevance theorists agree that the gap between the semantic representation of sentences and the thoughts communicated by utterances can be filled by inference, but not more coding; which similarly corresponds to the results of the ontology mapping surveys discussed in Section 1.2. They point out that semantics can only help to determine the possibilities of interpretation. "*An utterance that explicitly expresses one thought may implicitly convey others*" [Sperber and Wilson, 1995]. "Explicitly" is related to the semantic representation of

the sentence uttered, which is constraint, whereas those that are implicitly contained are free of constraints. Pragmatics from a relevance theoretic point of view is a supplement to grammar, and as that it substantiates the code model of verbal communication [Carston, 1998].

"*Communication is successful not when the hearers recognize linguistic meaning of the utterance, but when they infer the speaker's meaning from it*" [Sperber and Wilson, 1995]. Analogous Fetzer [2004] points out that "*meaning is mainly built up from the context of use*". Pragmatics is an investigation of *meaning in context* or *meaning in use* [Carston, 1998]. Therefore, pragmatics is characterized as the study of contextual factors in verbal communication. Sperber and Wilson [1995] indicate that "*a speaker who intends an utterance to be interpreted in a particular way must also expect the hearer to be able to supply a context which allows that interpretation to be recovered*"; and they further point out that "*a mismatch between the context envisaged by the speaker and the one actually used by the hearer may result in misunderstanding*", akin to the pragmatic heterogeneity problem in ontology alignment (cf. Section 1.2). A way for being "on the safe side" that such misunderstanding does not occur would be to make sure that both contexts are identical—the one of the speaker and that of the hearer. That requires properties of context to substantiate communication properties.

In the framework of the code model *mutual knowledge* is assumed for this purpose. The stage for mutual knowledge between communicator and audience has to be set by a common purpose. Grice [1975] calls such a purpose a *Cooperative Principle*, which has to be at least a "*mutually accepted direction*". As a result context has to be strictly limited to that mutual or common knowledge. Relevance theorists argue that people may look at the same object and yet identify it differently; they may impose different interpretations on information, and they may fail to recognize facts [Sperber and Wilson, 1995]. Therefore, there is no guarantee in assuming mutual knowledge. Carston [1998] criticizes that determining which knowledge communicator and hearer share and which not would have to perform an infinite series of checks and that would be a highly time-consuming, cost-intensive process. Relevance theorists hold that mutual knowledge is a construct "*with no close counterpart in reality*" [Sperber and Wilson, 1995]. Speech act theorists differentiate between the coding-decoding mode and the inferential mode; whereas relevance theorists use the inferential process as part of the decoding process. A more detailed insight in the two theories of pragmatics and their representatives is given for instance by Bach [2002], Carston [1998], Sperber and Wilson [1995].

### Relevance Theory

The *relevance theory (RT)*, introduced by Sperber and Wilson [1995], is considered as *contemporary pragmatics*, which is an interdisciplinary field in contrast to classical pragmatics. Relevance theorists consider two layers of communicative behavior: a basic layer of information, which is exploitable by the code model theory, and a second layer consisting of the evidence that the first layer has been made manifest. In the theory of Sperber and Wilson [1995] relevant information "*improves an overall representation of the world*". The goal of this theory is to consider *contextual effects* of explicitly expressed assumptions.

> "*Contextual effects is a necessary condition for relevance, and that other things being equal, the greater the contextual effects, the greater the relevance*" [Sperber and Wilson, 1995].

In their approach relevance is a *comparative and quantitative (C&P)* concept,—a property of mental processes by which "*the ordinary notion of relevance*" [Sperber and Wilson, 1995] of an utterance can be defined. They take as background the *Logical Foundations of Probability* introduced by Carnap [1950]. He presents in his theory three types of concepts:

1. *Classificatory concept* for classifying things into two kinds (e.g., warm or cold, big or small). This concept can be a property or relation.
2. *Comparative concept* is a relation based on comparison, e.g., some things are higher than others.
3. *Quantitative concept* for numerical comparison, e.g., distance, temperature, price. This kind of concept may correspond to the classificatory concept (e.g., temperature corresponds to the property warm).

"*Comparative concepts serve for the formulation of the result of a comparison in the form of a more-less-statement without the use of numerical values*" [Carnap, 1950]. The comparative concept stands between the two other kinds. Objects can be easily compared by formulating absolute judgements [Sperber and Wilson, 1995]. Studies in the field of *cognitive engineering* [Norman and Draper, 1986] have shown that human cognition is guided by *relevance*. Carston [1998] state more precisely: "*the human cognitive system is oriented towards the maximisation of relevance*". The expert's knowledge is context dependent, personally constructed, and highly functional [Agnew et al., 1993]. A *non-logic functional* or cognitive view of domain objects facilitates judgements and comparisons of intentions [Sperber and Wilson, 1995]. For the realization of such a view *propositional attitudes* [Carston, 1998] are necessary by which such judgements can be facilitated. A propositional attitude is a relation between the communicator and their uttered proposition. It captures the communicator's knowledge of the intended meaning of the proposition. The relevance of an assumption is what the communicator intends it to be in a certain context. This means that the communicator determines the relevance of an explicitly expressed assumption.

In verbal communication *ostensive behavior* (or *ostension*) provides evidence of the communicator's thought [Sperber and Wilson, 1995]. Ostensive acts are public, and therefore observable. For instance: The communicator says something and skeptically raises the eyebrow. The goal is to make manifest an intention (e.g., by eyebrow-raising) for making something manifest (e.g., (s)he means business) to the hearer [Carston, 2008]. Raising the eyebrow is an evidence that there is some relevant information to be obtained. In verbal communication such non-coded "*ostensive behaviour provides evidence of one's thought; it implies a guarantee of relevance*" [Carston, 1998]. Relevance theorists equate inferential communication (cf. Secion 4.4) and ostension. They view them as the same process. In that form of *informative communication* the communicator is involved in ostension and the audience in inference [Sperber and Wilson, 1995]. An *informative intention* is defined as "*to make manifest a set of assumptions to the audience*" [Sperber and Wilson, 1995]. The communicator must have a representation of such a set in mind

when uttering an assumption. Generally, the goal of informative intention is to help focusing the attention of the audience on relevant information, because the audience neither is able to decode, nor to deduce the communicator's intentions. "*The best (s)he can do is construct an assumption on the basis of the evidence provided by the communicator's ostensive behaviour*" [Sperber and Wilson, 1995].

A crucial factor in human interaction is "*maximising the relevance of information processed*" [Sperber and Wilson, 1995], which is a "*cognitive principle of relevance*". Firstly, discussed by Grice [1975] who distinguishes in his work between *saying* (semantics) and *implicating* (pragmatics). He was interested in the separation of "*what our words say from what we, in uttering them, imply*" [Grice, 1975]. What is said is truth-conditional, whereas what is implicated is non-truth-conditional. In his work he concerns on "beyond" what is said. The core of his studies are *conversational implicatures*, which indicate him as the founder of inferential pragmatics [Lab, 2006]. He assumes that the conversational implicatures are grounded in common knowledge of what the speaker has said. They depend on a set of maxims concerning the presentation of information. He denotes that maxims as *cooperative principles*, which concern:

- *Quantity:* make your contribution as informative as required, but make it not more informative as required;
- *Quality:* try to make your contribution one that is true;
- *Relation:* be relevant;
- *Manner:* avoid obscurity expression; avoid ambiguity; be brief; be orderly [Grice, 1975].

Sperber and Wilson [1995] criticize that these maxims are norms or rules, which communicator and audience must know and agree in order to communicate adequately. They state that "*implicatures are explained as assumptions that the audience must make to preserve the idea that the speaker has obeyed the maxims, or at least the co-operative principle*" [Sperber and Wilson, 1995].

Relevance theorists view the principle of relevance as intended for explaining ostensive communication explicitly and implicitly. They assume that people have *intuitions of relevance*. Such intuitions make it feasible to distinguish relevant from irrelevant, or less relevant information. The focus on relevance as comparative concept facilitates *intuitive judgements of relevance*, which are suggestive and not conclusive [Sperber and Wilson, 1995]. This means that contextual effects are not the same as contextual implications, but there is a very close connection between them; viz. if something is relevant, because it is related to context then it yields a contextual implication [Sperber and Wilson, 1995]. The authors use relevance as evidence for the hearer in the inferential process of verbal communication (cf. Section 4.4). The authors define relevance as classificatory concept expressed in necessary and sufficient conditions;

> "*An assumption is relevant in a context if and only if it has some contextual effect in that context*" [Sperber and Wilson, 1995].

The second factor of RT, which is involved in achieving contextual effects is the *processing effort*. Sperber and Wilson [1995] state that "*contextual effects are brought about by mental*

*processes*", which indicate a "*certain expenditure of energy*". The processing effort is unlike contextual effects a negative factor. There is a tradeoff between these two factors: "*the greater the processing effort, the lower the relevance*" [Sperber and Wilson, 1995]. At this stage the authors extend their approach by defining relevance as *comparative concept*. The two extent conditions, which imply necessary and sufficient conditions are;

> "*Extent condition 1: an assumption is relevant in a context to the extent that its contextual effects in this context are large;*"
> "*Extent condition 2: an assumption is relevant in a context to the extent that the effort required to process it in this context is small*" [Sperber and Wilson, 1995].

So far there are no quantitative values involved by which contextual effects and processing effort could be compared. Sperber and Wilson [1995] point out that "*contextual effects and processing effort are non-representational dimensions of mental processes*". The authors hold that these two factors are represented in the form of intuitive *comparative judgements*. They comment that people can take advantage of their comparative abilities in trying to maximize the relevance of information they process. Further, the authors state that relevance is a function of effect and effort, and as such a non-representative property. "*Relevance is a property which needs not be represented, let alone be computed, in order to be achieved*" [Sperber and Wilson, 1995]. If relevance is represented, then it is represented in the form of comparative judgements (e.g., irrelevant, weakly, relevant) instead of quantitative ones.

Classical pragmatists assume that context is generally given. Context is seen as uniquely defined. The representatives of RT act on the assumption that there is more than one context on hand to humans; there exists a range of possible contexts. "*The selection of a particular context out of that range is determined by the search for relevance*" [Sperber and Wilson, 1995]. This assumption leads to the following definition:

> "*Relevance to an individual: an assumption is relevant to an individual at a given time if and only if it is relevant in one or more of the contexts accessible to that individual at that time*" [Sperber and Wilson, 1995].

Gruber [1995] discusses that one of the major criteria in designing an ontology is *clarity*, which means that the intended meaning should be effectively communicated. Therefore, we focus on pragmatics from a relevance-theoretic point of view to enrich the semantic representation with informal, cognitive meta-information. Our aim is to enhance the intended meaning interpretation in ontology alignment in order to meet the requirement of "clarity".

## 4.5 Cornerstone of CoMetO: the Modeling Focus

Pragmatists distinguish between explicitly communicated assumptions (*explicatures*), and implicitly conveyed assumptions, or *implicatures*, which have been not communicated and are therefore not explicit. This corresponds to a comment made by Uschold et al. [1998] in the course of their investigation of ontology reuse for aircraft design. They state that there are "*hidden assumptions that were implicit in the original code of the ontology*", which makes them

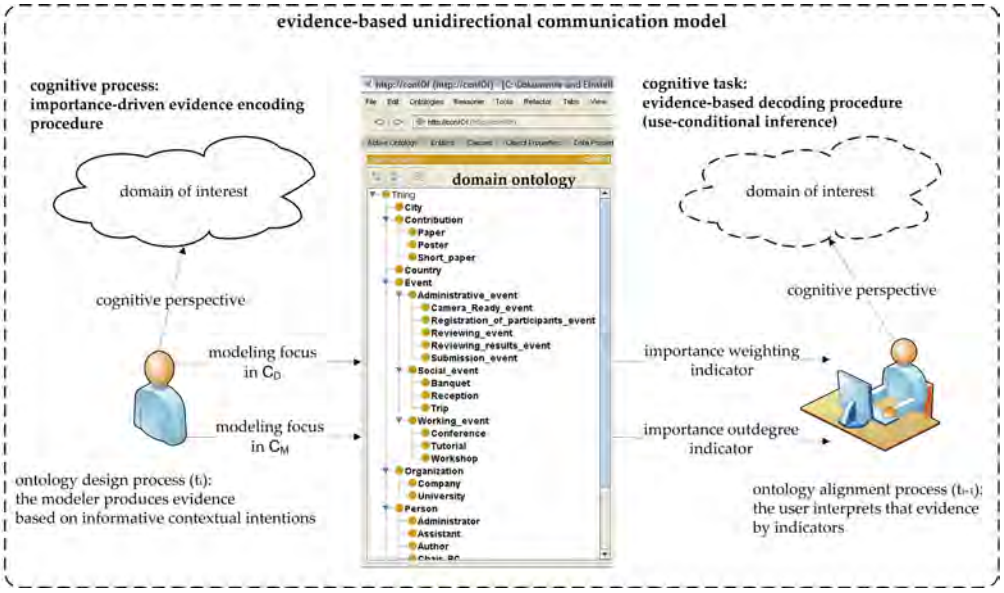



Figure 4.2: Meaning negotiation from the modeler to the user.

private and unobservable. We consider ontology development and alignment as to be comparable with the classical, pragmatic-based code model. Corresponding to Definition 3 (cf. Section 4.3) modelers make a set of assumptions explicit by using $\mathcal{L}$ and maybe guided by formal competency questions (conceptual encoding) and model-based techniques decode the explicitly defined meaning (semantics) in ontology alignment. This means that methods based on model theory bring two code models into mutual agreement, thereby semantic interoperability is provided. We take a step forward by presenting the idea of implementing a system akin to the relevance-based inferential model of verbal communication (cf. Section 4.4) in order to make it feasible for modelers to give information of the implicitly defined meaning (pragmatics) to users. Figure 4.2 illustrates this approach. The graphical example shows an *evidence-based unidirectional communication* model as "intermediate" construct between an ontology's design ($t_i$) and its alignment process ($t_{i+1}$), which temporally diverge. The left side of the graphic shows the theoretical part of CoMetO introduced in this chapter, and the right side shows the methodological part that is presented in Chapter 6. In this model we consider implicitly made assumptions as *informative contextual intentions* that are based on the modeler's cognitive perspective on the domain. Our approach is that modelers annotate the (explicitly) expressed assumptions with context-based meta-information when developing an ontology (importance-driven evidence encoding procedure) and users are able to interpret those evidence, corresponding to the modeler's intention, in ontology alignment (evidence-based decoding procedure). By implementing this model we provide a basis for a use-conditional form of meaning consideration as well as interpretation.

People are able to share *cognitive environments* because they can share physical ones [Carston, 1998]. Their shared cognitive environment is said to be a *mutual cognitive environment* [Sperber and Wilson, 1995] that is created by inferential communication, which is based on

ostensive acts from the communicator to the audience (cf. example of "eyebrow-raising", Section 4.4). In CoMetO we consider the domain of interest as the physical environment a modeler, as well as a user keep in mind when performing tasks in ontology engineering and alignment. In order to implement a shared (mutual) cognitive environment, we introduce a concept by which firstly on a theoretical level the modeler's cognitive perspective on domain concepts related to certain contexts ($C_D$, $C_M$) can be represented. We name this concept the *modeling focus (MF)*. Sperber and Wilson [1995] define as cognitive environment a function of the physical environment and an individual's cognitive perception. For this purpose we form the concept of MF as a component of the modeler's design process knowledge by which informative contextual intentions (cognitive semantics) are linked to schema level entities (i.e., logical statements), by which domain concepts are described.

There are three layers in the context of reusing ontologies (cf. Figure 2.1 in Section 2.3). Using the classification introduced by Ehrig et al. [2004] there is a horizontal dimension including: data, ontology, and context layer; and a vertical dimension—the domain knowledge layer, "*which can be suited at any layer of the horizontal dimension*" [Euzenat and Shvaiko, 2007]. We relate MF to the domain knowledge layer that makes it feasible for us to assign it to the ontology (model) layer. This is an initial stage for a solution to consider pragmatics (i.e., context-based semantics) at the model layer, as required by Janiesch [2010] (cf. Section 2.2).

We consider MF as an evidence of the modeler's intentional view of domain concepts and their relationship related to $C_D$ and $C_M$ when developing an ontology from scratch. Norman and Draper [1986] define an intention as "*the decision to act so as to achieve the design goal*". This means in our approach that MF reflects the modeler's intention to act so as to satisfy the ontology's design goals. Thus, it constitutes *expert meta-knowledge* that we classify as follows:

**Definition 4.**

$$\textit{modeling focus} \sqsubseteq \textit{process knowledge} \sqsubseteq \textit{background knowledge}$$

The modeling focus is a component of the modeler's process knowledge at design time ($t_i$), which is as such a component of domain-related background knowledge when aligning ontologies ($t_{i+1}$).

The *modus operandi* of CoMetO is that modelers make their context-based intentions explicit as informal expressions at the ontology's schema level. Producing such evidence by linking the modeling focus to logical statements is a *cognitive process* that—akin to the pragmatic-based relevance theory—is driven by relevance, but unlike this theory the mental properties in our approach are quantifiable and representational (which makes them visible to users). In order to avoid misunderstandings we use the term "importance" instead of "relevance" in CoMetO. After this process the modeling focus is encoded in two parameters: the *importance weighting indicator* and the *importance outdegree indicator*, which we present in the methodological part introduced in Section 6.3. In this encoded form the engineer's *modeling focus* is an evidence of domain-related background knowledge and as such it is a cognitive aid as recommended in Section 1.2. The parameters indicate the importance of classes compared to other classes related to their context-based usage (in $C_D$ as well as in $C_M$) at the schema level.

In the works presented in Section 2.2 researchers consider context as a single environment in which the model viz. its concepts have a certain meaning akin to the metaphysical theories discussed by Bouquet et al. [2002] (cf. Section 2.3). In our approach we understand context as a part of the cognitive perspective, corresponding to the cognitive theories [Bouquet et al., 2002]. This means that we consider the ontology, the modelers' view of the domain, and contexts in a single environment. The difference of our approach to other context-based ones is that if context is considered as an environment that stands for its own, similarly to the ontology itself, then a separate layer viz. a context layer is necessary for its consideration in ontology alignment (cf. Section 2.3). Additionally, we do not have to imitate the design process of an ontology by introducing *ex post* knowledge as presented in the approach of Wu et al. [2008] and other works as outlined in Section 2.2.

## 4.6 Concluding Remarks

"*The syntactic level interfaces with the internal conceptual system*" [Chomsky, 1980], and the semantic level with the external conceptual system, but "*there is more to be considered than is obvious at first thought*" [Norman and Draper, 1986]. Therefore, we introduced a cognitively inspired design methodology—CoMetO. The main issues of this methodology are: (i) we view *model-theoretic semantics* as the relation between the logical form of expressed statements (sentence meaning) and the entities in the real world; and (ii) *cognitive semantics* as the relation between that statements and their meaning in a certain context (sentence meaning in context), as intended by the modeler; (iii) we adapt *pragmatics* viz. its *relevance theory* in order to make cognitive semantics visible to users when aligning ontologies.

We stated that it is important to distinguish between the modelers' logical view in combination with the ontology language, and their cognitive view of the domain (cf. Section 4.3). The latter subjective view facilitates the analysis of the intentional context [Benerecetti et al., 2001]. For its consideration we presented the idea of a relevance-theoretic system that interfaces with the use-conditional system of an artifact. For this purpose, we adapted the relevance-based inferential communication model of ostensive behavior (cf. Section 4.4). We introduced the concept of a *modeling focus* that is a component of process knowledge of ontology design, which is as such a component of domain-related background knowledge. Our aim is to improve model-based reasoning methods by an additional importance-driven evidence-based decoding system and not to replace them.

Another aim is to store that expert meta-knowledge at the schema level to provide a single, consistent environment for meaning interpretation, as proposed by Falconer et al. [2006], by which a kind of "traceable history" of design decisions is made feasible, as demanded by Park and Woo [2007]. The interpretation of such specified knowledge in the form of *context-based indicators* is made practicable by a formal model, which we introduce in the following chapter. For this purpose, we extend OWL DL to provide a mechanism by which cognitive (context-based) constraints on statements are facilitated in order to populate an ontology's structure with pragmatic-specific instances.

CHAPTER 5

# CoMetO Metamodel

In this chapter we present a formal model—the *CoMetO metamodel*, in order to implement the theoretical part of our approach where we discussed to consider the modeling focus as relation between the logical form and the modeler's cognitive perspective on schema level entities. The implemented metamodel provide an *importance-driven evidence encoding procedure* by which such a linkage is made feasible to foster a use-conditional form of meaning consideration. For this purpose we define an extension to the OWL DL metamodel by adding *Meta Object Facility* [OMG, 2006] constructs for populating the ontology layer with specific instances. The language constructs of OWL DL that we use as input are the `owl:ObjectProperty`, the `rdfs:domain` and the `rdfs:range` axioms, which we supplement with *cognitive constraints*. The meaning of these basic elements from a logical perspective is already defined, which guarantees a shared interpretation; this means that an ontology captures *consensual knowledge* accepted by a group or community [Gašević et al., 2006]. Concepts and their relationship in the domain are in set theory classes and relations that constitutes the structure of the ontology. We focus on classes and their relations to other classes—together—as a subset, which can be annotated with *weightings* by the original modelers. Such an (informal) expression represents a public making act similar to ostensive acts in verbal communication (cf. Section 4.4).

## 5.1 Conceptual Design

Generally, the used techniques implemented in alignment tools reconstruct domain-specific knowledge based on an analysis of semantics, the schema, or data of the sources (cf. Section 2.1). We propose an alignment support for users already integrated in the ontology's design process and conducted by the original engineers themselves (cf. Section 4.2). They use language constructs to formally represent knowledge (logical perspective); what is missing is to pragmatically express knowledge (cognitive perspective), which is a non-logical form of knowledge representation. We fill this gap by our approach of considering the usage of ontology entities in certain contexts (cf. Section 4.5), whose implementation we describe in this chapter, and whose practical application we present in Chapter 6.

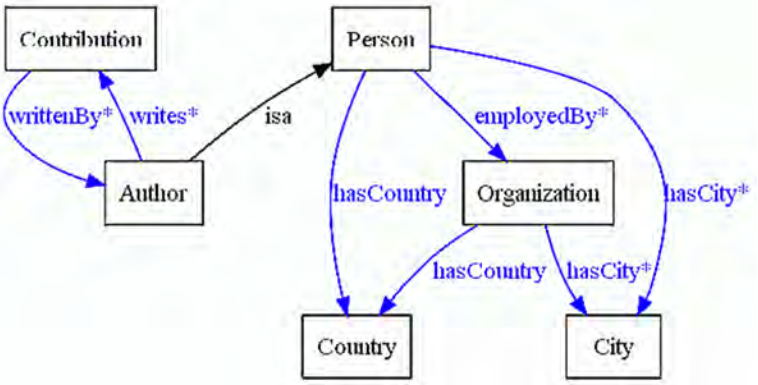



Figure 5.1: Sample of the `confOf` ontology visualized as labeled directed graph via OntoViz.

In our approach we focus on the schema level entities (at the ontology layer) and omit instance data (cf. Section 4.3). The ontology layer can be sub-classified in four levels: (1) *semantic nets*, (2) *description logics*, (3) *restrictions*, and (4) *rules* [Ehrig, 2007]. We relate the CoMetO metamodel to the lowest level of that layer,—the *semantic network*. At this level an ontology can be viewed as a labeled graph. Figure 5.1 illustrates a part of the `confOf` ontology (cf. Section 3.3) as such a graph visualized by using OntoViz[46] a TabWidget implemented in the Protégé editor. The classes are represented as nodes or vertices and the relations among them as directed labeled edges or arcs. Formally, the graph of an ontology's schema is a structure that can be defined as follows:

$$G = (V,\ A,\ \sum_V,\ \sum_A,\ s,\ t,\ l_V,\ l_A)$$

where

- $V$ is a set of vertices and $A$ is a set of arcs;
- $\sum_V$ and $\sum_A$ are finite sets of vertex and arc labels expressed in natural language. $\sum_V$ is a set of single- and multi-words (e.g., *Author*, *Administrative_event*), and $\sum_A$ is a set of verb phrases (e.g., *hasCity*, *writes*);
- $s : A \rightarrow V$ and $t : A \rightarrow V$ are two functions indicating the source and target vertices of arcs;
- $l_V : V \rightarrow \sum_V$ and $l_A : A \rightarrow \sum_A$ are two functions (i.e., vertex labeling, arc labeling) by which the vertices and arcs of G are mapped to a finite set of labels ($\sum_V$, $\sum_A$).

There may exist more than a single binary relation between two classes, since there are many relations among concepts in a domain. Therefore, the graph structure constitutes a *labeled multigraph* [Euzenat and Shvaiko, 2007]. A multigraph is a *digraph* (directed graph) with multiple arcs (parallel directed edges), i.e., arcs with the same source and target nodes [Matoušek

[46] A visual browser for ontologies, `http://ontoviz.sourceforge.net/` (last accessed August-1-2011).

and Nešetřil, 2007]. Thus, two nodes may be connected by more than one arc, for example[47] (`crs_dr` ontology):

$$\begin{array}{lcl} \textit{has\_author} & \rightarrow & \textit{(article, person)} \\ \textit{has\_reviewer} & \rightarrow & \textit{(article, person)} \end{array}$$

There also may exist cycles within that graph. In graph theory this means that an arc links a node to itself, for example (`confOf` ontology):

$$\textit{follows} \quad \rightarrow \quad \textit{(Administrative\_event, Administrative\_event)}$$

In this section we prefer the term "node" as used in common literature for labeled vertex, the term "arc" instead of "directed labeled edge", and for shortening the term "graph" instead of "labeled multigraph".

As stated in many contributions [Bouquet et al., 2003b, Euzenat and Shvaiko, 2007, Magnini et al., 2003, Noy and Musen, 2001], to the subject of "context consideration by the graph structure of ontologies", the meaning of a node depends not only on its label, but also on the position of the node in the graph. Commonly, researchers use WordNet in order to interpret the meaning of a node's label, as well as the meaning of the nodes' labels in the neighborhood of that node for context interpretation. For instance, in the contribution of Magnini et al. [2003] (cf. Section 2.2) they use WordNet in order to code the description of a label's meaning in description logics. In CoMetO we also consider the context of a node's (local) neighborhood as in other works, but with the crucial difference that we include expert meta-knowledge of the original modelers.

Euzenat and Shvaiko [2007] introduce three types of a graph structure: (1) the *taxonomic structure*, (2) the *mereologic structure*, and (3) the *relational structure*. The taxonomic structure considers `rdfs:subClassOf` relations that constitute the hierarchy of a graph, whereas the mereologic structure contains `part-of` relations among classes [Euzenat and Shvaiko, 2007]. We assume that context-based knowledge is mainly represented in the relational structure of an ontology and not necessarily in its hierarchical structure (taxonomy), since "*the world is complex, and a hierarchy is often too simple to capture the essence of the relationship between things*" [Passin, 2004]. A taxonomy describes rather containment relationships [Rahm and Bernstein, 2001] with, which we assume, only marginal inferable contextual evidence. Therefore, we focus on that structure where classes are related through the definitions of their properties when implementing our approach.

Based on the idea of *weighted graphs*, where semantic content can be communicated in the form of weighting the nodes and (directed) edges of a graph [Ottmann and Widmayer, 2002], we implement a metamodel by which the ontology's (relational) structure can be enriched with cognitive semantics. For this purpose, we provide a basis to link the modeling focus in $C_D$ and in $C_M$ by *importance weightings* (*iweightings*) to each `owl:ObjectProperty` (arc) with its certain `rdfs:domain` (node) and `rdfs:range` (node) axioms, which we consider—together—as a single schema level element (i.e., a proposition). By these weightings experts give information of the entities' usage in terms of their "importance" in the domain description to users. The `rdfs:domain` defines the kind of things the object property may apply to, whereas

[47] We use the formal notation as defined by Ehrig [2007] for an ontology's relational structure.

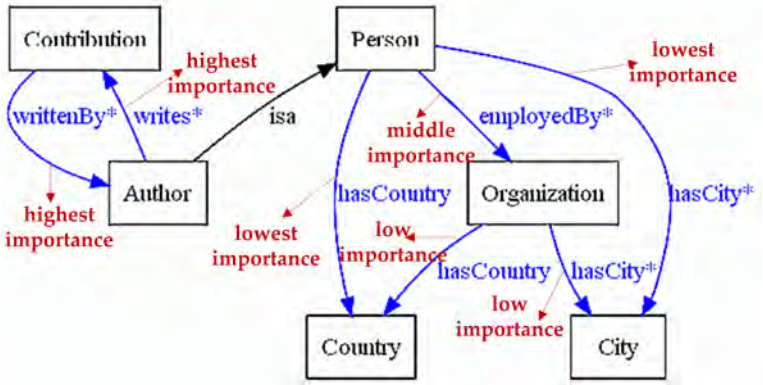



Figure 5.2: Importance weighted relational graph structure.

the `rdfs:range` defines the values the property is allowed to take (cf. Section 3.2). These axioms constitute logical constraints for semantic-based reasoning (cf. Section 4.3), while the iweightings are cognitive constraints (pragmatic- and structure-based constraints) and have as such no effects on model-based semantics. We use CoMetO to join the modeler's logical and cognitive perspective on the domain. Figure 5.2 illustrates the first procedure of *pragmatic-based constraints* by annotating each binary relation with an iweighting depending on its domain/range axioms and on the importance of that proposition in the domain context (i.e., its contextual effect). Figure 5.3 presents the second procedure of *structure-based constraints* where each node (labeled domain class) is annotated with the number of its outgoing relations to adjacent nodes within the schema. In graph theory that number constitutes the *outdegree* $d^+(v)$ of a vertex. We consider the outdegree of each node as its local context when implementing the method for computing indicators for users in ontology alignment (cf. Figure 4.2 in Section 4.5).

Object properties induce the *relation signature* $\sigma$ of a graph [Ehrig et al., 2004]. Ehrig [2007] formally defines such a signature as a function;

$$\sigma : R \rightarrow C \times C, \text{ where } \sigma(r) = \langle dom(r), ran(r) \rangle \text{ with } r \in R$$

where binary relations are interpreted as a subset of the *Cartesian product*[48] of the sets of two classes. For example:

$$\sigma(writes) = (Author, Contribution)$$

In OWL DL such relations are modeled by the `owl:ObjectProperty` construct (cf. Section 3.2). Domain/range restrictions on object properties are fulfilled by class expressions, which are individuals (i.e., resources) residing at the data layer.

[48]The Cartesian product or Cross product ($A \times B$) is a set of all possible ordered pairs ($a$, $b$); in that $a$ is an element of set $A$ and $b$ is an element of set $B$ [Matoušek and Nešetřil, 2007].

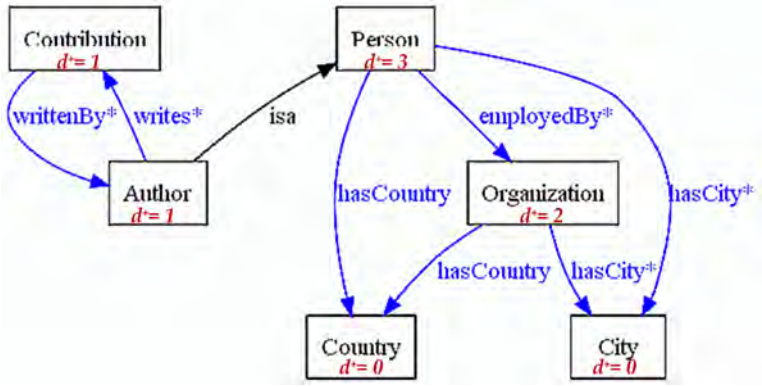



Figure 5.3: Outdegree $d^+(n)$ per node.

**Definition 5. Importance weighting**
We define an importance weighted relation signature $\sigma(R)_{iw}$ as a quadruple;

$$\sigma(R)_{iw} = \langle \sigma,\ l_v,\ l_a,\ \omega \rangle$$

where $\sigma$ is the signature function as defined before; $l_v$ is a function that maps the domain and range classes (i.e., nodes) to a set of natural language-based labels; $l_a$ maps the binary relations to a set of arc labels; and $\omega$ is a weighting function by which elements of $\sigma$ can be annotated with elements of $IW$, $\omega :\ (R \rightarrow C \times C) \rightarrow IW$; $IW$ is a finite set of importance weighting labels, $IW$ = {highest importance, high importance, middle importance, low importance, lowest importance}.

For example;

$$\omega(\sigma) :\ (writes \rightarrow (Author,\ Contribution)) \mapsto \texttt{highest importance}$$

this means that the proposition *writes* $\rightarrow$ (*Author*, *Contribution*) is annotated by a pragmatic-based constraint in the form of an iweighting label "`highest importance`". In CoMetO we exclusively consider the relation signature of the ontology, i.e., its *relational structure* [Euzenat and Shvaiko, 2007], which represents the set of all statements formally expressed at the model layer.

OWL facilitates to express metalevel information about entities in the form of *annotation properties*. There exist several predefined properties (e.g., `rdfs:label`, `rdfs:comment`, `owl:versionInfo`), and modelers can define further ones. The annotations can be used to associate additional information in the form of metadata with ontologies, entities, and axioms [W3C, 2004], but in a restricted way when using OWL DL to the effect that the filler (`AnnotationPropertyValue`) of such properties must either be a data literal, an IRI[49], or an individual, and they have no domain range sets [W3C, 2004]. It is not possible in OWL DL to annotate a logical statement at the ontology layer with meta-information, as it can be performed

[49]IRI = Internationalized Resource Identifier

by the *reification* of statements at the data layer when using the language RDF [W3C, 1999]. For instance, a certain statement (at the data layer) made about the resource "Dr. Smith" can be written in RDF/XML syntax as follows:

```
<Author rdf:ID="Dr.Smith">
 <writes rdf:resource="#ABC"/>
</Author>
```

RDF provides a mechanism called reification for making (higher-order) statements about statements at the data layer. For this purpose the original statement is made to a model, which is then a new resource called *reified statement* to which additional properties can be attached [W3C, 1999]. For instance:

```
<rdf:Description>
   <rdf:subject resource="Dr.Smith" />
   <rdf:predicate resource="writes" />
   <rdf:object>ABC</rdf:object>
   <rdf:type resource="Statement" />
   <a:date>2010-07-12</a:date>
 </rdf:Description>
```

In the example above we attach another property `date` to the reified statement with the filler "`2010-07-12`". By doing so we express that a researcher, Dr. Smith, has written an article (*ABC*) on a given date. Generally, such a construct can be used as an object of other statements, or they can have additional statements on it [W3C, 1999]. The basic idea is to generate $n$-ary (multi-valued) relations, but when using OWL DL (as well as OWL2) only binary relations are supported.

We extend OWL DL to overcome that limitation. The annotation we practice is comparable to that of a propositional attitude in pragmatic theory, which constitutes the relation between a communicator and their uttered proposition (cf. Section 4.4). The importance weightings consider non-logical information (similar to annotation properties), and therefore they induce no semantic conflicts. Although we extend that language in the version of 2004 [W3C, 2004], our model can be equally applied to the most recent version (OWL2 as of 2009). Firstly, when we started the work on this thesis the current version was in draft. Secondly, we use the elements `owl:ObjectProperty`, `rdfs:domain`, and `rdfs:range`. In OWL2 [W3C, 2009] these schema level entities are neither modified, nor there is a new construct presented by which the meta-annotation of a proposition, or a sentence at the model layer is facilitated. Therefore, it makes no difference which one we use, because both versions have to be extended for the implementation of the CoMetO metamodel. Moreover, tool support (e.g., ontology editors) is much better for the already established version 2004 than for the relatively new OWL2 variant. The language's extension facilitates to store the meta-knowledge within an ontology in order to provide a single environment in the alignment process.

## 5.2 Conceptual Modeling

A domain ontology describes a particular domain of interest (cf. Section 3.1). It constitutes a full specification of that domain [Gašević et al., 2006]. In OWL DL we distinguish between the ABox, which contains *assertional knowledge*, and the TBox, which contains the *terminological knowledge* of a domain and its intances [Hitzler et al., 2008]. The TBox includes all essential concepts, their description, their classification, as well as the relations that hold among them [Rahm and Bernstein, 2001], while the ABox contains the data axioms (i.e., the individuals of that concepts and the statements which they are belonging to). Thus, the TBox defines semantic relations among individuals that are instantiated in the ABox. The TBox is assigned to the model-structure (ontology) layer and the ABox to the information (data) layer according to the four-layer metamodeling architecture introduced by the Object Management Group (OMG[50]). Figure 5.4 illustrates the levels (M0-M3) of that architecture and presents the dependencies by an example. We use UML [51] [Jeckle et al., 2003] and its stereotype notation for visualizing a part of the `confOf` domain ontology. UML® and the *Meta Object Facility (MOF™)* provides a key foundation for the OMG's *Model-Driven Architecture (MDA®)* [OMG, 2009].

The lowest level (M0) is the data model (i.e., information layer) where facts about individuals of the domain are stated. For example: Dr. Smith writes a contribution entitled "ABC".

```
<Author rdf:ID="Dr.Smith">
    <writes rdf:resource="#ABC"/>
</Author>
```

The rules for creating facts about individuals are determined by the schema (at the ontology layer) the data model is related to. At the conceptual (M1) level the schema of a certain domain is residing. The labels of elements, formal-based semantics, as well as the content rules (i.e., syntax rules) are defined in such a schema [Rahm and Bernstein, 2001]. For example, $writes \rightarrow (Author,\ Contribution)$ expressed in OWL DL:

```
<owl:ObjectProperty rdf:about="#writes">
     <rdfs:domain rdf:resource="#Author"/>
     <rdfs:range rdf:resource="#Contribution"/>
</owl:ObjectProperty>
```

A domain-specific metadata schema is expressed in a certain schema definition language (e.g., OWL DL). That language is described in the model at level M2. In this metamodel the terms and operators for building expressions are defined. It reflects the language primitives (abstract syntax), a concrete notation (concrete syntax), and semantics by which the schema is constrained (e.g., domain/range construct) [Haslhofer and Klas, 2010]. In OWL DL these specifications are defined in description logics (DL), which makes the semantics of language constructs to decidable fragments of first-order logic [Gašević et al., 2006]. M3 contains a universal (self-defining) modeling language,—(E)MOF by which metamodels can be specified, constructed, and managed [OMG, 2009]. There is a direct dependency between the levels (M0, M1, M2,

[50] `http://www.omg.org` (last accessed August-3-2011).

[51] UML = Unified Modeling Language

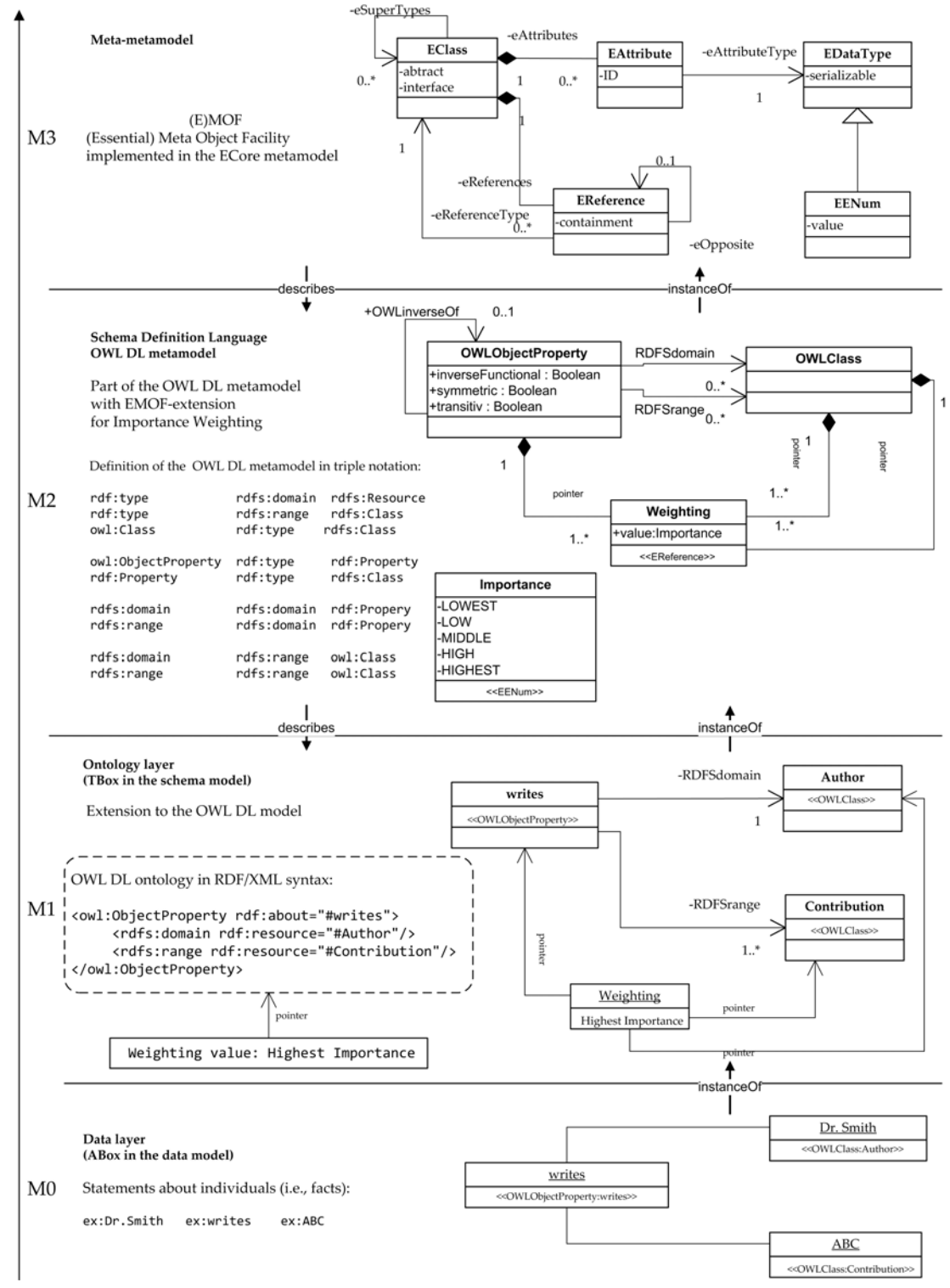



Figure 5.4: The four-layer Model-Driven Architecture.

and M3) denoted by the `instanceOf` relation in Figure 5.4. Such a linear structure makes it feasible to make assertions in M2 of instances in M1.

We extend parts of the OWL DL model by a formal model by which entities at M1 can be supplemented with (informal) pragmatic- as well as structure-specific instances. Thereby, we distinguish between logical (model-based) constraints, which are global (e.g., domain/range axioms); and cognitive constraints, which are local and context-dependent. The distinction between those two categories is given in that logical constraints supports model-based reasoning, which is truth-conditional; while cognitive constraints facilitate use-conditional inference made by the user. The UML diagrams at the right side of the levels M1 and M2 (cf. Figure 5.4) illustrate such an extension using the structural features of the ECore metamodel at M3. ECore is a meta-metamodel similar to (E)MOF (they are parallel modeling spaces), although it is more simple and implementation-friendly [Gašević et al., 2006].

## 5.3 Extension to the OWL DL Metamodel using EMF

We use the *Eclipse Modeling Framework (EMF)* [Budinsky et al., 2004] for our purpose. We leverage constructs of EMF for defining new elements at the M2-layer in order to instantiate elements at M1 by using the *ECore metamodel* [Budinsky et al., 2004]. ECore is the core metamodel in EMF, which supports the main concepts of the Model-Driven Architecture (MDA). MDA provides a framework for defining domain-specific languages on the basis of MOF [OMG, 2006]. MOF is a key standard in the MDA family, as it is the basis of the OMG´s MDA. EMF is an open source model-driven software development platform and an efficient Java[52] implementation of a core subset of the MOF API. Using the concepts of MOF we are able to define the abstract syntax of schema definition languages (meta-languages). MOF has two parts: *Essential MOF (EMOF)* and *Complete MOF (CMOF)* [OMG, 2006]. For the implementation of our approach we use EMOF, which provides a straightforward framework for metamodels, and makes it feasible to define that models by using simple concepts. EMOF is implemented in the EMF´s ECore metamodel. Therefore, EMOF is compatible with ECore and it can be used to define and extend a meta-language as OWL DL.

For the realization of the importance weighting approach we use the ECore class `EReference` as basis. This class is a kind of pointer by which the ends of a binary relation can be represented. Thereby, a certain object property with its domain/range constraints can be viewed as a single information unit, which can be annotated by an importance weighting label (pragmatic-based constraint) as intended by the modeler. The `EEnumerator` data type facilitates to represent the weighting degree for the `EReference` class `Weighting` by using literals. Additionally, to the introduced structural features ECore provides a construct to model the behavioral features of a class as `EOperations`. The bodies of operations must be coded by hand in the generated Java class. Figure 5.5 illustrates the extended OWL DL metamodel by using the constructs of the ECore metamodel at level M2. The source code is reproduced in the Appendix (A.1-A.3).

The granularity level of the CoMetO metamodel is that of individual elements. Therefore, we only define parts of an OWL ontology by using the *Ontology Definition Metamodel*

[52] `http://www.java.com/de/download/` (last accessed August-3-2011).

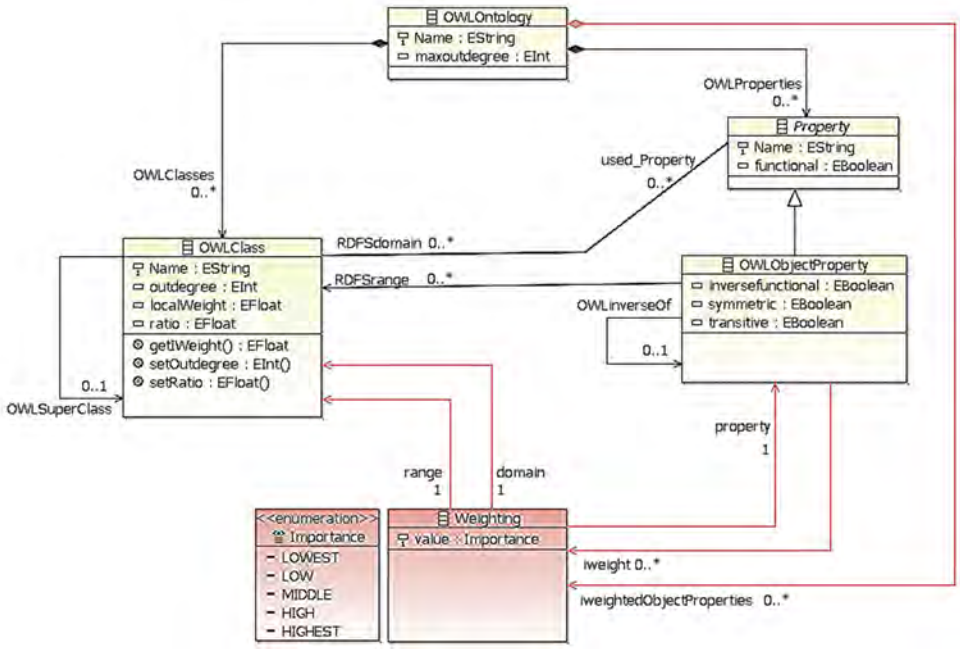



Figure 5.5: The CoMetO metamodel.

*(ODM)*. ODM classes could be seen as MOF instances [OMG, 2009]. Thereby, we use EMOF as abstract syntax for the extended CoMetO metamodel. The extensions, `Weighting` and `Importance`, are highlighted in order to distinguish them from the elements of the ODM-based OWL metamodel. The `owl:Ontology` is defined by `OWLOntology` and described by the attributes `name` with the data type `string` (e.g., *confOf*, *crs_dr*), and `maxoutdegree`, which is an `integer` that outputs the highest outdegree of a class within the ontology's schema. An `owl:Class` is defined as `OWLClass` with its attributes: `name`, `outdegree`, `localWeight`, and `ratio`. It may have a super class (`OWLSuperClass`). We implement an algorithm, which we denote as `getLocalWeight` (cf. Appendix A.1), that computes the manually annotated importance weightings (cf. Figure 5.2) into a numerical value for each (domain) class; a procedure which we introduce in detail in Section 6.3. The attribute `LocalWeight` outputs that computed value, which constitutes the *pragmatic-based constraints*. The `setOutdegree`-algorithm (cf. Appendix A.2) adds the used object properties of classes that constitute their `outdegree`, as illustrated in Figure 5.3. The `setRatio`-algorithm (cf. Appendix A.3), which bases on the classes' outdegree and the `maxoutdegree` within the ontology is also presented in detail in Section 6.3. The attribute `ratio` outputs that computation as numerical value, which constitutes the *structure-based constraints*.

The presented constructs aided us to implement a prototype of the CoMetO metamodel, which we name the *Odm_ExtensionModelEditor (OdmEMEditor)*. That editor is implemented

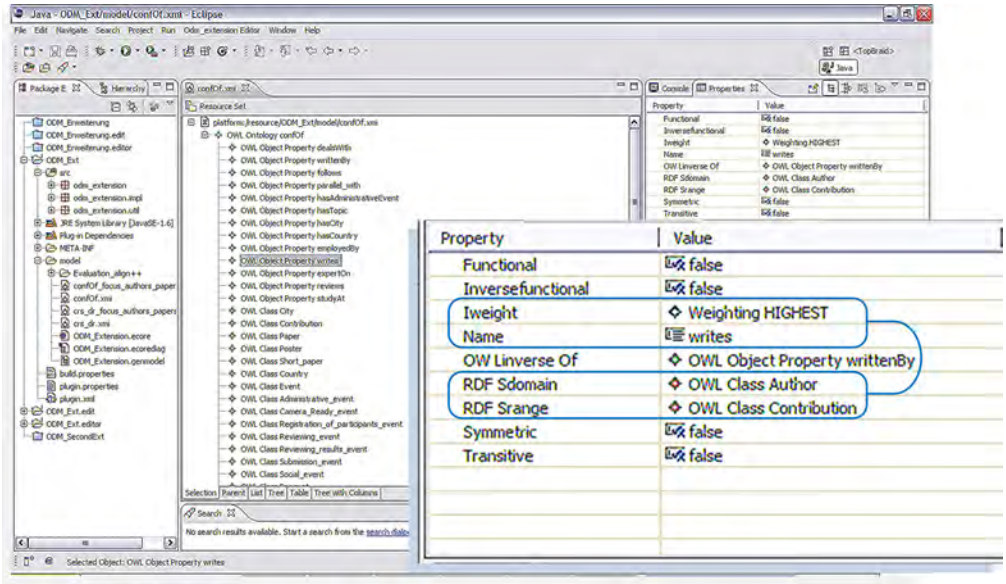



Figure 5.6: Example of an iweighting annotation to a proposition of the `confOf` ontology in the OdmEMEditor.

in Eclipse in the version Galileo[53]. Figure 5.6 shows the model in its practical application by weighting the proposition *writes* $\rightarrow$ (*Author*, *Contribution*) of the `confOf` example ontology with the iweighting label `highest importance`. Figure 5.7 illustrates the editor window presenting the outdegree of the domain class (*Author*) in that logical statement with a value of 1 ($d^{+}(\texttt{Author}) = 1$).

## 5.4 Concluding Remarks

We took an MOF-space point of view by leveraging the constructs of ECore for defining and integrating a formally, well-grounded metamodel—the CoMetO metamodel. We extended OWL DL by using simple class modeling concepts in order to enrich the relational structure of an ontology with the capability to cope with our approach. We integrated constructs of ECore (`EReference`, `EEnumerator`, `EOperation`) in order to populate the schema model at the ontology layer with specific instances, which constitute cognitive constraints.

The extension of OWL DL is similar to the approach introduced by Vrandečić et al. [2006] with the crucial difference that we did not create an external metamodel as proposed in their work. In CoMetO we do not differentiate between a domain and its context-based information as distinct universe of discourses (i.e., physically separated ontologies). Thus, we do not require complex formalisms to bring that knowledge into controlled interaction, e.g., by bride rules [Bouquet et al., 2002, 2003a, Giunchiglia, 1992, Giunchiglia and Serafini, 1994]. On one hand,

[53] `http://www.eclipse.org/downloads/packages/release/galileo/sr2` (last accessed February-1-2010)

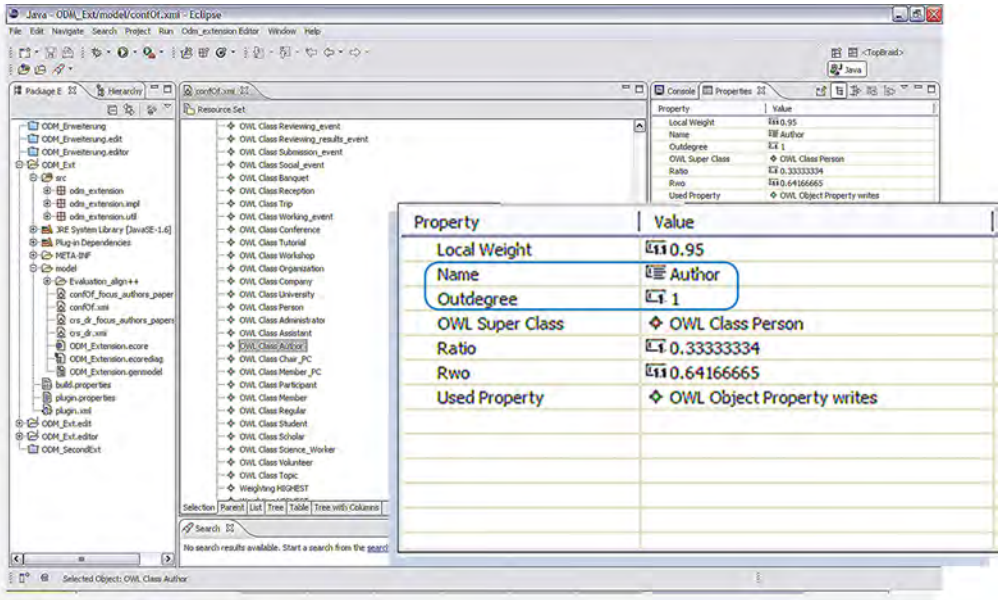



Figure 5.7: Example of an outdegree annotation to a class of the `confOf` ontology in the OdmEMEditor.

researchers describe metalevel information as corresponding to the domain; on the other hand they create new ontologies, or meta-views to deal with this kind of information, which leads to independent interpretations (e.g., where queries are used to integrate information). In our approach we differentiate between two forms of interpreting meaning: a logical or model-based form, and a cognitive, pragmatic-based one.

"*Cognitive science is the study of human thinking in terms of representational structures in the mind, and computational procedures that operate on those structures*" [Hofstadter, 1995]; we present both issues by the methodological part of CoMetO, which we introduce in the following chapter. We demonstrate the usefulness of the CoMetO approach in practice on examples based on the `confOf` and `crs_dr` domain ontologies. We supplement these sources by annotating their relational structure with pragmatic- and structure-based constraints based on nontrivial application scenarios. We use the OdmEMEditor in order to perform the computations of a method that users can benefit from, prior to initiating a schema-based alignment.

CHAPTER 6

# Methodological Part of CoMetO

In this chapter we present the methodological part of our approach, where we use the implemented constructs of the CoMetO metamodel. We introduce a "bridging method" by which an *unidirectional evidence-based communication* from the modeler to the user is facilitated; the idea of which we developed in Section 4.5 (cf. Figure 4.2). We name this procedural method "*align++*". It is based on a heuristic, since most cognitive processes are too complex that they can only be approximated. The name "align++" results from the two parts of which this method consists; an *ex ante Part A* where importance-driven evidence encoding procedures are conducted by the modeler, and an *ex post Part B* where the user is aided in the evidence-based decoding task that is performed prior to initiating an alignment. The first step of Part A involves two *weighting methods*: a *direct* and an *indirect* one. In a second step of this part algorithms compute *contextual parameters*. These parameters support users when interpreting meaning in a use-conditional form. We blur the distinction between linguistic-based and structure-based techniques with the procedures introduced in Part A. In Part B of align++ we use concepts of inferential and financial statistics for the implementation of two *mismatch-at-risk metrics*. These metrics are based on probability theory. They make it feasible to additionally interpret the computed metadata of Part A with regard to possible heterogeneities, which may occur when aligning ontologies. We follow the objective to aid users for a better understanding of the sources prior to conducting an alignment process in order to disburden them from cognitively complex, time-, and cost-intensive tasks.

## 6.1 Part A of align++

Schema-based mapping methods analyze mainly two factors: entity labels and relations among entities (cf. Section 2.1). The aim of schema-based methods is "*to guess the meaning encoded in the schemas*" [Shvaiko and Euzenat, 2004]. As we pointed out in previous chapters there is a semantic (explicitly defined) meaning and a pragmatic (implicitly defined) meaning. The latter is based on the intended use of schema level elements. Therefore, we proposed that beside

the entities' labels and their position in the schema a factor of *importance*, resulting from their contextual effect, should be additionally considered. We introduced two contexts in which entities are used: the *domain context* ($C_D$) and the *modeling context* ($C_M$) (cf. Section 4.1). The entities' usage in one of these contexts is based on the original modeler's cognitive perspective. This perspective can be represented by the introduced concept of a *modeling focus (MF)* (cf. Section 4.5). This focus constitutes a relation between the logical form of statements (sentence meaning) and the engineer's cognitive perspective when expressing them (sentence meaning in context).

In the first part of align++ we present such a linkage, which is made by cognitive constraints in the form of *importance weighting annotations*, on the basis of a practical example. In that process we interpret a logical statement as a phrase in some natural language related to the context in which it is processed. We introduce algorithms by which *contextual parameters* are computed based on the constraints. These parameters are automatically mapped onto classes, additionally to their labels, in the form of two indicators: an *importance weighting indicator* and an *importance outdegree indicator*. By presenting these indicators we make it feasible to enrich the element level with additional meta-knowledge of the ontology's schema level; to counter the fact that considering only one level leads to information loss coded in the other. The annotated meta-information is recorded in the schema itself in order to have a single, consistent environment as input in the alignment process. Our aim is to provide a representation that is rich enough to describe semantics as expressive as possible in a model- and cognitive-based form.

## 6.2 Importance-driven Evidence Encoding

The encoding procedure is a *cognitive process* where *importance* is a mental property; similar to the comparative and quantitative concept of the relevance theory (cf. Section 4.4). Our aim is that modelers can express their informative contextual intentions (cognitive semantics) in order to make them visible to users; in a similar way as a communicator makes informative communicative intentions explicit to an audience (cf. Section 4.4). The public making act in our approach are weighting annotations, which are based on importance-driven comparative (mental) judgements. Such a mental process is conducted by the modeler through *direct* and *indirect importance weightings*. Firstly, in a semi-automated step on logical statements; and secondly, in an automatically computed step on classes in the role of a `rdfs:domain` in that statements. Such iweighting annotations constitute informal meta-information that does not effect the logical aspects of the ontology, which are needed for reasoning over a set of given facts (cf. Definition 3 in Section 4.3). Our aim is to foster a non-logical improvement of ontology alignment as proposed by researchers discussed in Section 1.2.

### Direct Importance Weighting Procedure

The *direct (semi-automated) weighting procedure* is manually conducted through the modeler by attributing each logical statement ($s_i \in S$, $i = 1, \ldots, n$) with an *importance weighting (iweighting) label*. A statement's ($s_i$) weighting degree bases on the importance of $s_i$ in $C_D$. Based on Sperber and Wilson's definition of "contextual effects" (cf. Section 4.4) we hold that

| Importance Weighting Label | Description |
|---|---|
| `highest importance` | The proposition has a highest contextual effect in $C_D$. |
| `high importance` | The proposition has a high contextual effect in $C_D$. |
| `middle importance` | The proposition has a middle contextual effect in $C_D$. |
| `low importance` | The proposition has a low contextual effect in $C_D$. |
| `lowest importance` | The proposition has a lowest contextual effect in $C_D$. |

Table 6.1: Importance weighting degrees

the greater the contextual effect the greater the importance of that statement in $C_D$. In our approach the modeler determines the degree (e.g., `highest`, `high`, `middle`, `low`, `lowest`) of importance a statement has compared to others and their contextual effects at the schema level (cf. Table 6.1). This means that unlike to the relevance theory the modeler evaluates a statement's meaning compared to others and makes that quantified judgement explicit. More formally expressed, we define:

**Definition 6. Importance weighting degree**
If $\langle iw_{lowest},\ iw_{low},\ iw_{middle},\ iw_{high},\ iw_{highest} \rangle$ is a scale of importance ($iw_{lowest}$ = the lowest importance, $iw_{highest}$ = the highest importance) then attributing statement $s_i$ with $iw_a$ and statement $s_j$ with $iw_b$ implies that $s_j$ has more contextual effect than $s_i$; $\forall\ iw_a,\ iw_b \in\ IW,\ iw_a < iw_b$.

The pragmatic-based condition "having a contextual effect to some degree" means in our approach that concepts are often classified as more or less important concerning the fulfillment of the purpose-specific design goals of an ontology. That fact corresponds to the observations presented by Park and Woo [2007], Ramesh and Dhar [1992], Smart and Engelbrecht [2008], which we have described in previous chapters. The iweighting procedure requires non-trivial knowledge about the domain and its concepts related to the purpose to what they are modeled. The engineers can guide such a process by competency questions [Grüninger and Fox, 1995]. The importance of entities (their "degree of meaningfulness") is that what the modeler intends them to have in a certain context. The weighting procedure is based on a mental judgement, in the relevance theory-based sense, where the engineer distinguishes more important statements from less important ones. By the first part of align++ such a comparative judgement is made visible to users, similar to ostensive acts in verbal communication (cf. Section 4.4).

We assume that engineers prefer to assign importance labels instead of numerical values in order to affix cognitive constraints on logical statements. The iweighting annotation function can be carried out by a simple point-and-click interaction. The mock-up presented in Figure 6.1 shows a user-friendly mechanism for attributing an iweighting label on the `owl:ObjectProperty` (*writes_article*) and its particular domain (*author*) and range (*article*) axioms. We assume that such an annotation mechanism is practicable for modelers even if they develop large ontologies. We distinguish five iweighting labels corresponding to Carston

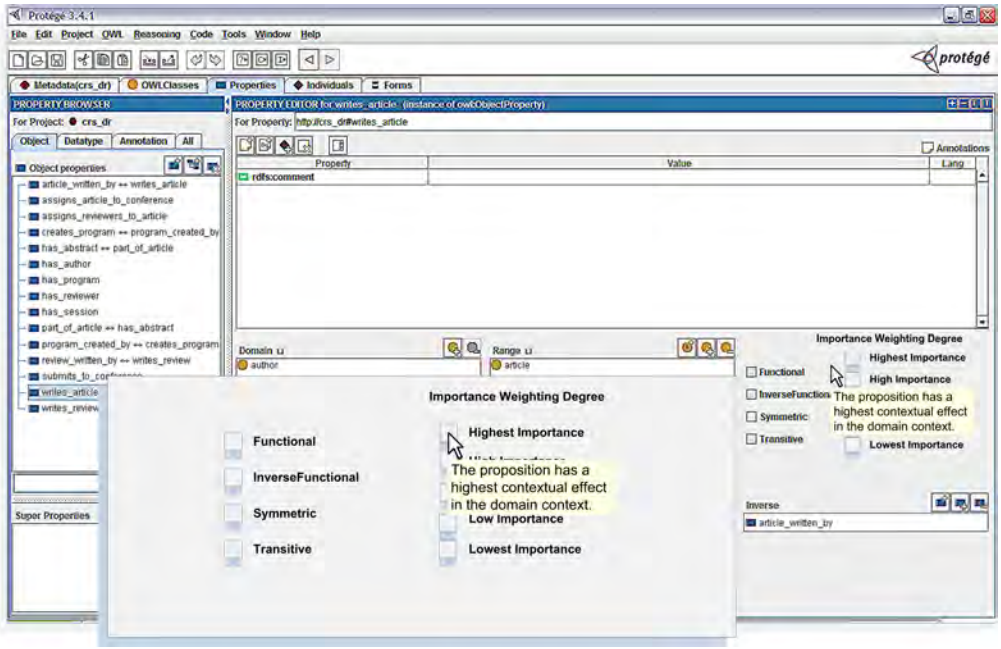



Figure 6.1: Example for implementing an iweighting annotation mechanism in the open source ontology editor Protégé in version 3.4.1.

[1998], who states that "*expectations of relevance may also vary in their specificity*". Table 6.1 shows that iweighting labels by which the importance of a statement's *meaning in use* can be expressed. Such an annotated label constitutes a pragmatic-based constraint on the semantic level of a logical statement. This means that the local semantics of statements are constrained based on their purpose-specific usage in context; that is no contradiction in Fetzer's view (cf. Section 2.3).

At this point we take an excerpt of the example ontologies $O_A$ and $O_B$ in order to gain a better understanding of what the engineer has to consider when annotating such informal constraints. We take the assumption that the design goal is to administrate authors and their submitted contributions. We evaluate the contextual effect of each statement compared to others as described in Definition 6. The issue of such effect can be guided by informal competency questions (cf. Section 4.3). They provide an initial evaluation of ontology entities by an informal justification, which may aid the modeler when determining the degree of a pragmatic-based constraint. For example:

$$s(1)_{iw} : \langle writes\_article(a, w) \rightarrow author(a) \wedge article(w) \rangle \mapsto \texttt{highest importance}$$

$$s(2)_{iw} : \langle article\_written\_by(v, b) \rightarrow article(v) \wedge author(b) \rangle \mapsto \texttt{highest importance}$$

$$s(3)_{iw} : \langle part\_of\_article(u, v) \rightarrow abstract(u) \wedge article(v) \rangle \mapsto \texttt{high importance}$$

$$s(4)_{iw} : \langle assigns\_article(c, p) \rightarrow author(c) \wedge conference(p) \rangle \mapsto \texttt{highest importance}$$

$s(5)_{iw} : \langle organizes(d, m) \rightarrow Chair(d) \wedge Event(m) \rangle \mapsto$ `low importance`

$s(6)_{iw} : \langle organized\_by(m, d) \rightarrow Event(m) \wedge Chair(d) \rangle \mapsto$ `lowest importance`

An alternative for allocating a weighting degree would be to count the number of questions that can be answered by the statement compared to others; the more questions can be answered the greater is the contextual effect. Another alternative would be to evaluate for each statement its significance of information for answering the questions. However, competency questions can guide ontology engineers in their cognitive process when producing contextual evidence for users. They are an aid when setting a suitable level of *iweight* to a proposition that fits its meaning in context. We assume that ontology engineers accept the recommendation proposed by Horridge et al. [2007] that an object property should have an inverse property (cf. Section 3.2). In the example, the object property *organizes* with *Chair* (domain) and *Event* (range) has not necessarily the same iweighting label as its inverse property *organized_by*; because the engineer determines the iweighting degree based on their focus in $C_D$.

### Indirect Importance Weighting Procedure

The *indirect (automated) weighting procedure* bases on the engineer's modeling style and is automatically computed by an algorithm at design time. Therefore, this procedure is an "indirect" cognitive process. The modeling focus in $C_M$ is based on the modeler's skills, preferences, and experience in modeling a domain. There are several options to use entities for expressing the domain's content. A concept can be described as a class, or as a qualifying attribute [Klein, 2001]. For example: engineers have two opportunities for making explicit the assumption that there exists a relationship between individuals of the concept "working event" to individuals of the concept "event". They can model this relationship as `rdfs:subClassOf` relation, expressing that each working event is an event;

$$Working\_event \sqsubseteq Event$$

or, by an `owl:ObjectProperty` ***hasEvent*** where this property links individuals of *Event* (`rdfs:domain`) to individuals of *Working_event* (`rdfs:range`);

$$\top \sqsubseteq \forall hasEvent^{-}.Event$$
$$\top \sqsubseteq \forall hasEvent.Working\_event$$

Different modeling styles may cause structural mismatch, which induces schema incompatibility (cf. Section 1.2). On the one hand, design decisions depend on the modeler's preferences; on the other hand, they result from the importance of a concept for the domain description. Noy and Musen [2001] point out that less important concepts should be described as an attribute instead of a class. As illustrated in Figure 3.3 (cf. Section 3.3) the more attributes are generated the more shallow is the structure of the ontology (cf. $O_B$), whereas the more classes are modeled the deeper is the hierarchical structure of the ontology (cf. $O_A$). The background knowledge of the modeling focus in $C_M$ is a useful evidence for users when initiating a graph-based alignment tool as Anchor-PROMPT (cf. Section 1.2). For this purpose we implement the `setOutdegree`-algorithm, which adds the number of outgoing relations of a class to other classes within the

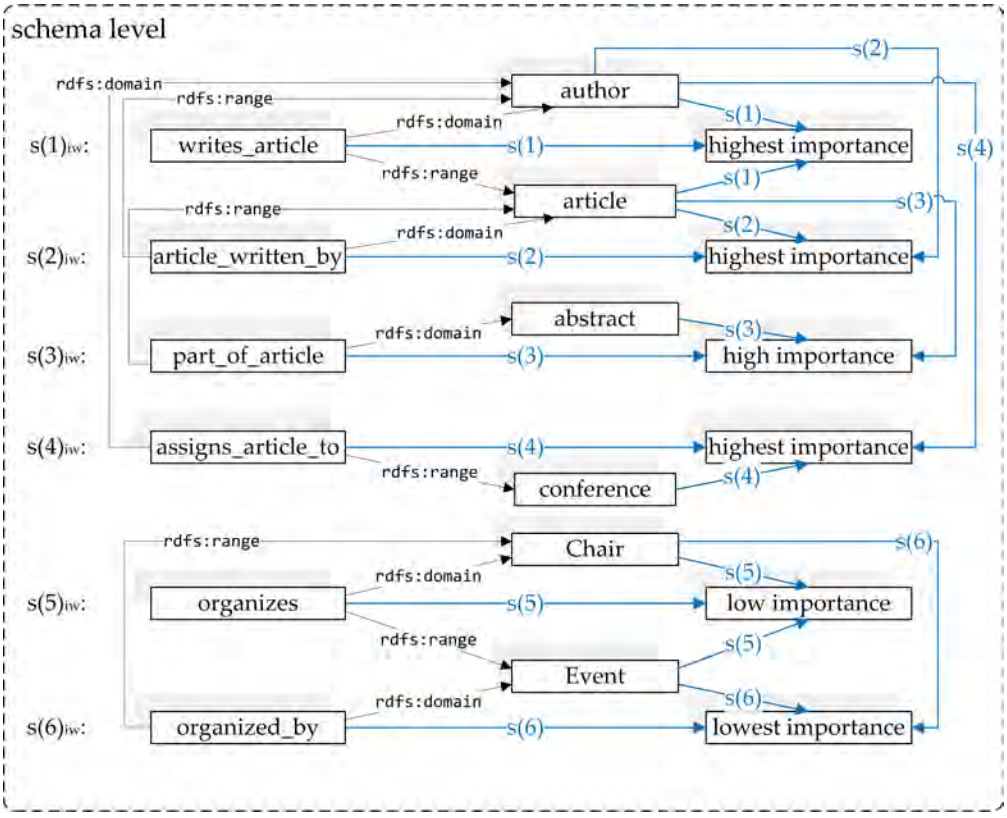



Figure 6.2: Example for recording the iweighted logical statements at the schema level (TBox) of the domain ontology.

schema for calculating a class' outdegree $d^+(c)$. This approach of an automatic "arc count" facilitates the approximation of schema differences on the basis of syntactical information.

An opportunity to record the importance weighting annotations is as key-value pairs in a multistage hash map. Figure 6.2 shows that for the manually (direct) conducted iweighting annotations of the example statements $(s(1)_{iw}, \ldots, s(6)_{iw})$. In the current version of align++, which is implemented by using the Eclipse environment (cf. Section 5.3), the direct and indirect iweightings are recorded in the form of ArrayLists[54].

We take the model-based semantic representation of the domain's syntactic logical form as input in the first part of the method of CoMetO. In an initial step *global* property restrictions as semantic axioms are supplemented by *local* pragmatic- and structure-based constraints in the form of iweighting annotations. The weights facilitate to tune the importance of semantics related to contexts ($C_D$, $C_M$); since "*certain mismatches between two expressions can only be detected from the context in which the expressions are used*" [Visser et al., 1997]. In a next step the annotated meta-information is automatically mapped to all classes in the role of a

[54] Java ArrayList, `http://leepoint.net/notes-java/data/collections/lists/arraylist.html` (last accessed June-11-2011).

`rdfs:domain` for computing contextual parameters.

## 6.3 Contextual Parameters

In this section we introduce calculation methods by which the cognitive constraints are converted to *contextual parameters*. These parameters contain the pragmatic- and structure-based constraints in a condensed, numerical form, firstly, to make them machine-readable; and secondly, to use them as input for further computations. The aim is to aid users in their decision-making process prior to initiate an alignment (e.g., for mismatch prediction). We want to overcome the obstacle that "*mappings cannot be defined beforehand, as they presuppose a complete understanding of the two conceptualizations, which in general is not the case*" [Bouquet and Serafini, 2000]. We focus on classes as the units of observation when computing such parameters. A class is an ontology element that has several features (e.g., label, constraints, restrictions, etc.) by which the concept of a domain can be described. We decide to enrich labeled classes instead of other ontology entities, because: (1) classes are information units by their particular linguistic notion. The content of a class is mainly described by its label, which is a language expression [Magnini et al., 2003]. Such a label functions as "*address in memory, a heading under which various types of information can be stored and retrieved*" [Sperber and Wilson, 1995]. (2) Labels are most important when identifying an alignment [Ehrig and Sure, 2005]. By computing contextual parameters as additional features of classes we meet the demand made by Shvaiko and Euzenat [2004] that "*in real-world applications, schemas/ontologies usually have both well defined and obscure labels (terms), and contexts they occur,* [sic!] *therefore, solutions from both problems would be mutually beneficial*". For this purpose we implement two algorithms (`getLocalWeight`, `setRatio`) by which the iweighted local context of each domain class is analyzed in detail and automatically mapped onto that classes in the form of two indicators: an *importance weighting indicator ($IwI_c$)* and an *importance outdegree indicator ($IoI_c$)*.

As introduced in our previous works [Mazak et al., 2010a,b] the interval-scaled $IwI_c \in [0,\ 1]$ importance weighting indicator is resulting from the direct iweighted semantics of binary relations, and the absolute frequency of the classes' role as `ObjectPropertyDomain` in that relations. The $IoI_c$ is ratio-scaled in the range over an interval of real numbers $[0,\ 1]$. This indicator is based on the indirect conducted iweighting annotations. The $IoI_c$ is a quotient resulting from the classes' outdegree in proportion to the highest outdegree within the ontology schema. Both parameters indicate the level of a class' importance compared to other classes in $C_D$ as well as in $C_M$ as intended by the modeler. Additionally, they can be used as estimators in order to approximate a structural and/or pragmatic mismatch between two ontologies prior to starting their alignment.

### Importance weighting Indicator ($IwI_c$)

The $IwI_c$ indicates a (labeled) class' importance in the domain description. For instance;

$$O_A : \textit{Author} \equiv O_B : \textit{author}$$

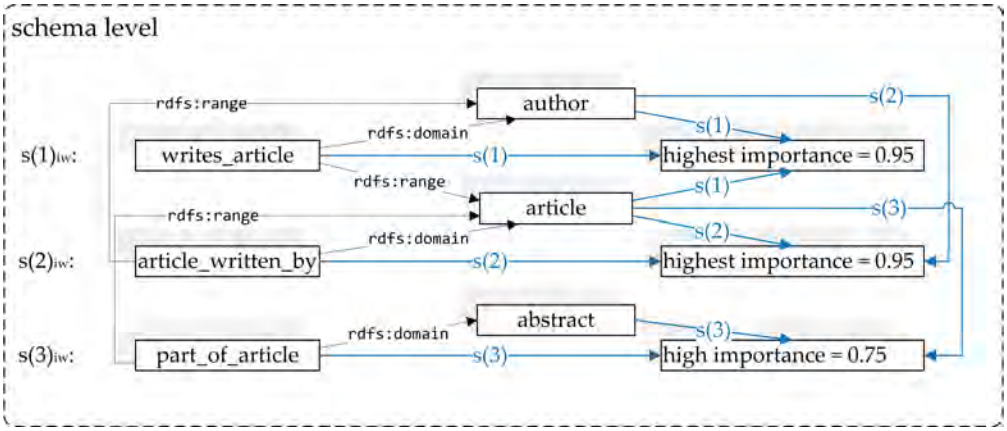



Figure 6.3: Conversion of the iweighting labels to numerical values.

implies not automatically that these two classes coincide in their information significance, too. For such a detection a use-conditional inference mechanism is needed in order to decode pragmatics. Such an exploitation would aid users to interpret classes (additionally to their labels) with contextual reference. We assume that significant information for such an inference process is rather involved in that statements where a class has a certain role (`rdfs:domain`), than in the class itself by considering it as a single information unit.

Before the $IwI_c$ of a class can be automatically computed it is necessary to convert the annotated iweigthing labels (cf. Table 6.1) to numerical values. We implement an *assignment function* $\psi$ in the `getLocalWeight`-algorithm (cf. Appendix A.1) for performing that task. We formally express the converting procedure, based on the notations introduced in Definition 5 (cf. Section 5.1), as follows:

**Definition 7. Numerical encoding procedure**
Let $\sigma(R)_{iw}$ be a finite or countable set of all iweighted propositions (i.e., `owl:ObjectProperties` with their `ObjectPropertyDomain` and `ObjectPropertyRange` axioms), which we consider as elements $\{\sigma(r_i)_{iw} \in \sigma(R)_{iw} \mid i = 1, \ldots, n\}$. A numerical encoding function $\psi$ is an injective mapping of these elements into a subset of $\mathbb{R}^+ = \{0.05,\ 0.25,\ 0.50,\ 0.75,\ 0.95\}$; $\psi : \sigma(R)_{iw} \rightarrow \{0.05,\ 0.25,\ 0.50,\ 0.75,\ 0.95\}$ based on a case distinction;

$$\psi(\sigma(r_i)_{iw}) = \begin{cases} 0.95 & \text{if } \sigma(r_i)_{iw} = \text{“}\texttt{highest importance}\text{”}; \\ 0.75 & \text{if } \sigma(r_i)_{iw} = \text{“}\texttt{high importance}\text{”}; \\ 0.50 & \text{if } \sigma(r_i)_{iw} = \text{“}\texttt{middle importance}\text{”}; \\ 0.25 & \text{if } \sigma(r_i)_{iw} = \text{“}\texttt{low importance}\text{”}; \\ 0.05 & \text{if } \sigma(r_i)_{iw} = \text{“}\texttt{lowest importance}\text{”}. \end{cases}$$

Figure 6.3 shows a part of the iweighted propositions, which were manually attributed in the example of Section 6.1, and are now converted to numerical values. In addition the directed

labeled ontology graph is now enriched with weightings. The procedure defined in Definition 7 is an automated mapping from the iweighting label space to a numerical value space.

**Definition 8. Importance Weighting Indicator ($IwI_c$)**
Let $c_i$ be a certain class $\{c_i \in C \mid i = 1, \ldots, n\}$ where $C$ is the set of all labeled classes ($l_V : V \rightarrow \sum_V$) within an ontology's schema; let $C_{Dom} \subseteq C$ be the set of all classes that take the role of a domain class; $\sigma(R_{c_i})_{iw}$ is a subset of manually iweighted logical statements where $dom(r)$ = $c_i \; \forall r \in R_{c_i},\; c_i \in C_{Dom}$; and $f_n(c_i)$ is the absolute frequency of $c_i$ as `rdfs:domain`. This implies that $\forall\, c_i \in C_{Dom} :\; f_n(c_i) > 0$ and $\forall\, c_i \in (C - C_{Dom}) :\; f_n(c_i) = 0$.

Given these definitions the `getLocalWeight`-algorithm aggregates the iweighted local context of $c_i$ and normalizes the sum by considering the absolute frequency of $c_i$ as `rdfs:domain` in the iweighted logical statements as follows:

$$IwI_{c_i} = \frac{1}{f_n(c_i)} \sum_{\sigma(r_j)_{iw} \in \sigma(R_{c_i})_{iw}} \psi(\sigma(r_j)_{iw}) \tag{6.1}$$

The normalization step makes the comparison to other classes feasible and is necessary for the computations of Part B. The higher the $IwI_c$-based value of a class the more importance has that class compared to other classes in the domain description.

Figure 6.4 and Figure 6.5 show in each case a pan of the implemented Odm_ExtensionModelEditor (cf. Section 5.3). The class *Author* of $O_A$ is related as `rdfs:domain` to one used object property (*writes*) and has a calculated $IwI_c$ of 0.95. The class *Person* of $O_A$ is related to three object properties: (1) *employedBy*, (2) *hasCity*, (3) *hasCountry*; and has a lowest $IwI_c$-based value of 0.20. The comparison between these two classes points out that the number of relations (*Author*:*Person* in proportion 1:3) is no indicator for a high importance of a particular class, as introduced in other works in Section 2.2 [Magnini et al., 2003, Wu et al., 2008].

## Importance outdegree Indicator ($IoI_c$)

Why do users need a second indicator of structural importance? Falconer and Storey [2007] state in their framework for cognitive support in ontology mapping that users need to understand the structural context of ontologies to obtain structural interoperability. Therefore, we need to make the engineer's modeling style visible to users that is represented by the focus in $C_M$. It cannot be derived from classes with a highest $IwI_c$ that they participate in many relations to other classes, which is an important fact when applying graph-based alignment techniques,–firstly, to detect efficient starting points, and secondly, to traverse as many paths as possible in the subgraphs (cf. Anchor-PROMPT in Section 1.2).

The importance of a class in $C_M$ can be quantified by its importance outdegree indicator ($IoI_c$). That parameter is automatically calculated by the `setRatio`-algorithm (cf. Appendix A.3) on the basis of a class' outgoing relations to other classes in proportion to the particular class with the most outgoing relations within the schema. The higher the $IoI_c$-based value of a class the more outgoing relations to other classes has that class, whereas $d^+(c) = 0$

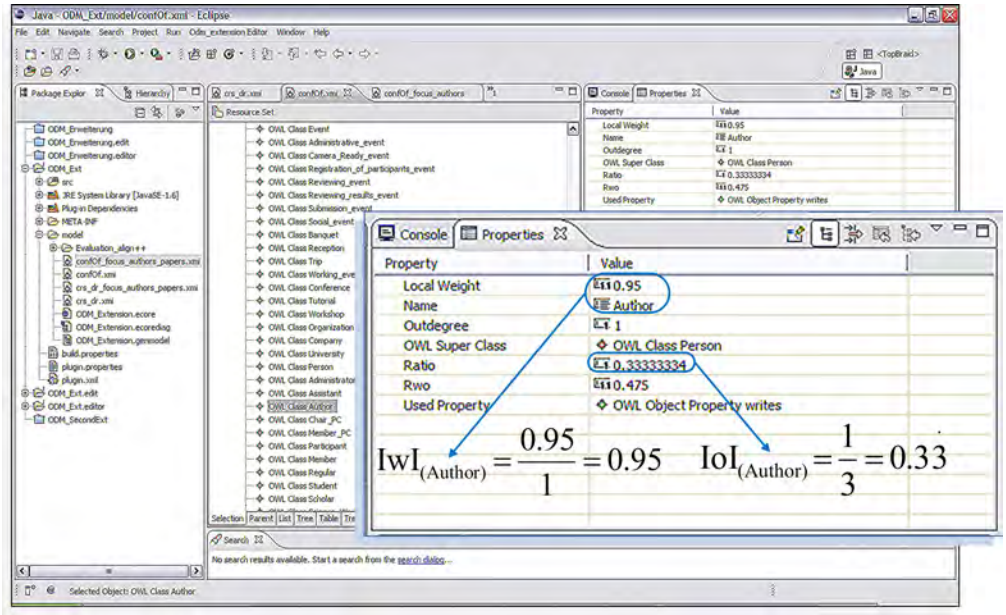



Figure 6.4: Example of the calculated $IwI_c$ and $IoI_c$ of the class *Author* of $O_A$.

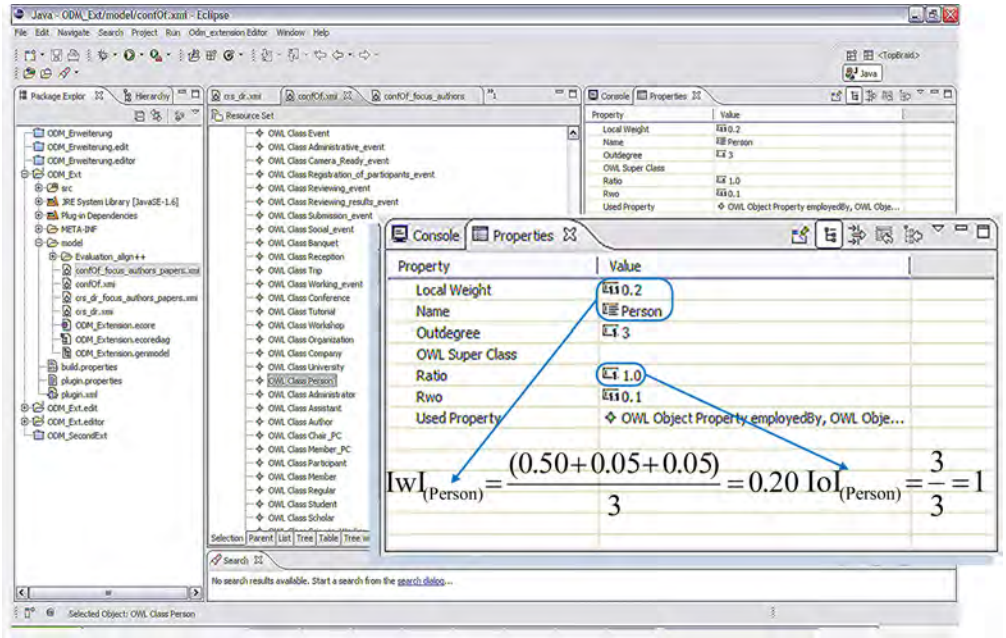



Figure 6.5: Example of the calculated $IwI_c$ and $IoI_c$ of the class *Person* of $O_A$.

is representing no outgoing relations. A low value indicates a class in a more taxonomic position, whereas a high value is an indicator for a network structure. This makes clear that there

is graph-based meta-information encoded in the classes' $IoI_c$-based values. More formally expressed, this means:

**Definition 9. Importance Outdegree Indicator ($IoI_c$)**
Let $c_i$ be a certain class $\{c_i \in C \mid i = 1, \ldots, n\}$ where $C$ is the set of all labeled classes ($l_V : V \rightarrow \sum_V$) within an ontology's schema; $d^+(c_i) = |a_j(c_i)|$ where $a_j \in A$ $(j = 0, \ldots, n-1)$ is a finite set of arcs where $c_i$ is the source node ($c_i : A \rightarrow V$); the $IoI_c$ of a class $c_i$ is calculated as the relative frequency of $c_i$ in proportion to the class with the maximum outdegree $max\ d^+(c_j)$, $c_j \in C$ (i.e., maximum number of relations where the class has the role of a domain) within the ontology's schema:

$$IoI_{c_i} = \frac{d^+(c_i)}{max_{c_j \in C_{Dom}} d^+(c_j)} \tag{6.2}$$

We revert to the pans representing each of which the calculated indicator-based values for the two classes of $O_A$ (cf. Figures 6.4, 6.5). The class *Author* participates in one relation to a different class, and has therefore a low outdegree in that example. This fact results in an $IoI_c$-based value of $0.3\dot{3}$ for *Author*, whereas *Person* with the most outgoing relations to other classes (i.e., maximum outdegree of 3) has a highest $IoI_c$ of 1. This example shows that *Author* is highly important in its meaning in use, but has only low importance, e.g., as a starting or terminating point for graph-based alignment. Quite contrary the class *Person*, which has a lowest importance in its implicitly defined meaning, but a highest importance outdegree.

In its first part align++ is a method by which expert meta-knowledge of an ontology's design process can be encoded and processed to indicators. The indicator-based values are metadata at the element (class) level, which are declaratively stored in repositories within an ontology. By the use of the introduced iweighting procedures and the subsequently conducted computations we meet the demand made by Euzenat and Valtchev [2004], who state that; "*to provide the most complete basis for comparison, one may wish to bring knowledge encoded in relation types to the object level*". The computed metadata are pragmatic- as well as structure-specific instances by which the ontology's schema is populated. This expert-dependent meta-information can be classified either to the functions such information is intended to support, or to the level of semantic abstraction (e.g., low level vs. high level metadata) [Benitez et al., 2001]. We classify the indicator-based values as higher-order metadata by which the internal resource of the modeler's cognitive perspective is represented in a machine processable and a user understandable form by which access to that knowledge is provided.

## 6.4 Evidence-based Decoding: Indicator-based Ranking Lists

Generally, the output of an alignment algorithm when comparing the entities of two domain ontologies are lists of candidates. Tools which provide such lists for users are for instance: FOAM [Ehrig and Sure, 2005], Chimaera and PROMPT. A problem is that such lists are difficult for users to understand and interpret, e.g., due to poor readability. This is one of the reasons why users need cognitive support in the interpretation task. We go even further by considering such a support prior to starting an alignment process. Mapping discovery is one of the major

| $IwI_c$ **Level** | **confOf** ($O_A$) | **crs_dr** ($O_B$) |
|---|---|---|
| *Highest* | Contribution | article |
| | Author | author |
| *High* | – | abstract |
| *Middle* | – | reviewer |
| | – | review |
| *Low* | – | – |
| *Lowest* | Administrative_event | conference |
| | Working_event | program |
| | Organization | chair |
| | Person | participant |
| | Member_PC | – |
| | Scholar | – |

(a) Scenario 1: equal modeling focus

| $IwI_c$ **Level** | **confOf** ($O_A$) | **crs_dr** ($O_B$) |
|---|---|---|
| *Highest* | Administrative_event | article |
| | Working_event | author |
| | Organization | |
| *High* | – | abstract |
| *Middle* | – | reviewer |
| | – | review |
| *Low* | Scholar | – |
| *Lowest* | Contribution | conference |
| | Person | program |
| | Author | chair |
| | Member_PC | participant |
| | – | – |
| | – | – |

(b) Scenario 2: different modeling focus

Table 6.2: $IwI_c$-grouped classes of $O_A$ and $O_B$ based on Sc. 1 and Sc. 2.

bottlenecks in ontology alignment [Noy, 2009] (cf. Section 1.1). Currently, there is no support of filtering large lists in order to categorize or group the candidates by certain characteristics [Falconer and Storey, 2007]. It is often difficult for users: (i) to get a quick and context-based overview of the sources; (ii) to know which concepts are the core concepts of those ontologies; (iii) to detect which concepts are good candidates as initial points for schema-based alignment techniques. Our aim is to disburden users from the need to analyze the structures of ontologies without efficient evidence.

There is a considerable amount of information derivable from the indicators. We assume that the re-usability of ontologies can be increased by both contextual parameters. They are an evidence for drawing the user's attention to potentially important concepts as intended by the original modeler. For instance, users can be guided when outlining an interesting sub-scope of the sources, e.g., for candidate selection. They can be used for ranking and grouping classes by their importance in the domain ontologies. That may help users to detect if the sources are structurally and pragmatically compatible, and which method (e.g., graph-based, model-based, or taxonomy-based) is better suited for aligning them. We implement two algorithms in align++ (`setRankIwI`, `setRankIoI`) for ranking the labeled classes by their calculated indicator-based values. The algorithms re-encode the classes' numerical $IwI_c$ and $IoI_c$ values, which are stored in ArrayLists as disjoint compositions into intervals (cf. Appendix A.4, A.5), by mapping these values to a lexical format for a better read- and understandability for users. After this procedure the ontology's domain classes are grouped in lists sorted in descending order. Such *ranking lists*, as presented in Table 6.2 and 6.4, are the output of the *ex ante* Part A of align++. They illustrate the modeling focus on classes in certain contexts ($C_D$, $C_M$).

Table 6.2 shows the lists in which the domain classes of $O_A$ and $O_B$ are ranked by their $IwI_c$-based values. The rankings are based on the modeling focus of the participants of our evaluation survey in each of the predefined application scenarios (cf. Section 7.2). The scores

are computed as average of the values resulting from the manually performed direct iweighting procedures (cf. Table 7.1). Those classes with a highest score are the core concepts. For instance, users can easily detect, using the lists presented in the left Table 6.2a, that the core concepts of the two ontologies are: *Author* of $O_A$ and *author* of $O_B$, which are also syntactically equal; and *Contribution* of $O_A$ and *article* of $O_B$. Additionally, the table shows that the lists may help users to take care of *terminological heterogeneity*, which occurs due to variations in names referring to the same concepts, like in case of *Contribution/article*. This means that both classes might be used to describe the same thing,—a written contribution to a conference. The two terms are used synonymously, but it is not straightforward to detect them as similar, neither by string-based techniques nor manually by users if they are not aware of the domain contexts. The lists depicted in the right Table 6.2b show differences in the classes' ranking compared to the lists in Table 6.2a. These differences are inferable due to the classes' dissimilar $IwI_c$-based values. It is evident that both ontologies describe the same domain of interest, but obviously with different modeling foci resulting from diverging design goals. Moreover, users can detect that the intended usage of the classes may differ. Thus, they can infer that aligning $O_A$ and $O_B$ may lead to pragmatic heterogeneity problems resulting in a mismatch between these sources.

Structural alignment methods require syntactically equal nodes (i.e., same labels) to use them as reference pairs for further mappings. Such pairs are defined either manually by the user, or automatically by lexical matching. For users it may be difficult to find useful sets of related terms especially if the sources are very large. Scores of matches can be found by lexical matching, but often they are not significant; for instance, *house* and *mouse* have a string similarity of $0.75$ computed by the edit distance. Generally, initial points are used to set new similar pairs by moving from one node to another via the directed edges among them.

To demonstrate the usefulness of comparing the classes' $IoI_c$-based values, we use Anchor-PROMPT (cf. Section 1.2) to align the example ontologies. The system suggests

- *Person* ($O_A$), *Event* ($O_A$); and
- *person* ($O_B$), *event* ($O_B$)

as initial pairs detected by lexical matching. By the `WalkPaths`-algorithm no correspondences can be found between these anchors, because there exist no relations between them. This fact could be easily detected when taking a look at the classes' $IoI_c$-based values prior to initiating the algorithm of this tool. Table 6.3 presents these values; only *Person* of $O_A$ has a highest outdegree and would be a good originating point for the `WalkPaths`-algorithm.

Table 6.4 shows the classes' grouping (high, middle, low) resulting from their $IoI_c$-based values. The user can infer that *Person* of $O_A$ and *author* of $O_B$, or *Contribution* and *article* are more efficient as anchor pairs. Each of these classes has a high or middle $IoI_c$ level, which indicates that they are involved in more than one relation to other classes within the schema. Actually, there exist two links between the classes *Person* and *Contribution* of $O_A$, and more than two links between *author* and *article* of $O_B$. This means that at least two paths of the subgraphs could be parallel traversed by the algorithm to detect more correspondences.

The CoMetO ranking lists are akin to the *repository of structure* technique [Euzenat and Shvaiko, 2007, Shvaiko and Euzenat, 2004] by which a comparison of ontology fragments is

| **class** | **domain ontology** | $IoI_c$**-based value** |
|---|---|---|
| *Person* | $O_A$ | 1 |
| *Event* | $O_A$ | 0 |
| *person* | $O_B$ | 0 |
| *event* | $O_B$ | 0 |

Table 6.3: Comparison of the $IoI_c$-based values among the detected classes of both ontologies.

| $IoI_c$ **Level** | **confOf** ($O_A$) | **crs_dr** ($O_B$) |
|---|---|---|
| *High* | Person | article |
| *Middle* | Contribution<br>Administrative_event<br>Working_event<br>Organization<br>Member_PC | author<br>program<br>chair |
| *Low* | Author<br>Scholar | abstract<br>reviewer<br>review<br>conference<br>participant |

Table 6.4: Ranking list of $O_A$ and $O_B$ resulting from the classes' $IoI_c$.

facilitated. This technique is used in schema-based alignment for first checking for similarity to the structures which are already available in the repository [Shvaiko and Euzenat, 2004]. The lists are a kind of *partial alignment* for candidate selection. They facilitate "alignment clustering", rather than discovering accurate correspondences among classes. They help users to feel confident that classes with equal labels also carry similar information significance. The ranking lists meet most of the requirements on cognitive aids as recommended by Falconer and Storey [2007] and Falconer et al. [2007] (cf. Section 1.2):

1. They require in their framework *lists of candidate mappings* which should support users by filtering classes, and by categorizing candidate mappings. By the indicator-based ranking lists users are aided, e.g., to filter the core concepts of the sources, and to identify efficient candidates for structure-based alignment methods.
2. The *identification of candidate-heavy ontology regions* is supported by reducing the complexity of the user's selection task by first identifying candidates with higher priority. The heuristic of align++ lowers the number of candidate mappings. Therefore, the lists are an efficient aid to prune the search space.
3. The consideration of the *context of mapping terms* is supported by taking into account the iweighted local contexts of each class in its role of a domain class. Thus, we are comparing

two classes with respect to their surrounding entities in the corresponding ontologies.

4. The *definitions for mapping terms*, which should include the properties of classes, and restrictions on those properties are satisfied by the modality of computing the direct ($IwI_c$) and indirect ($IoI_c$) weighting annotations.

5. The *inconsistency detection* where users should be supported to detect conflicts or inconsistency due to the candidates, is enhanced by Part B of the align++ method.

## 6.5 Part B of align++

A pertinent question that comes to mind when aligning two domain ontologies is the risk of a structural and/or pragmatic mismatch. Commonly, users are interested in those kinds of mismatch, because they both are currently unresolvable. The process to be examined in this *ex post* Part B of align++ is the approximation of such mismatches prior to initiating a schema-based alignment. The outcome of that *random process* is not certain. There is a trade-off between alignment chance and mismatch risk. In order to quantify the uncertainty of a possible mismatch in a real number, we use the construct of a *random variable* from probability theory. On the basis of that theoretical construct and a risk metric, which we adapt from financial statistics, we implement two heuristic-based *mismatch-at-risk* metrics. The (domain) classes' indicator-based metadata, computed in Part A, are the internal input on which both metrics operate. The statistical *variation*, which constitutes the spread among the classes' $IwI_c$-based values, is the *risk indicator* for approximating a pragmatic mismatch, whereas the risk drivers for a possible structural mismatch are the classes' outdegrees in relation to the total number of classes of each ontology. The risk metrics have both a heuristic nature, which means that exactness is sacrificed in favor of performance.

In the CoMetO approach we consider only schemas and omit instance data (cf. Section 4.3, Section 5.1). A schema is defined as a set of elements connected by some structure [Rahm and Bernstein, 2001]. Schema-based alignment techniques take as input two schemas and produce mappings among schema elements that correspond to each other. Alignment techniques detect similarities among entities of the input ontologies as: *equal*, *syntactically equal*, *similar*, *broader than*, *narrower than*, or *different* [Bouquet et al., 2004]. A user initiates an alignment process by specifying a source and a target ontology. Then the algorithm of a tool (e.g., Chimaera, or PROMPT) computes an initial set of candidate mappings in the form of a list largely based on lexical similarity of the classes' label. After that step the user works with this list to verify the recommendations, or to create mappings missed by the algorithm. Once the user has verified the mapping, the algorithm uses this anew to perform analysis. That usually results in further mapping suggestions and the process is repeated (cf. Section 1.1). Not until after that "longsome" user-guided alignment process mismatches can be detected, which are caused by heterogeneities among the sources. Such "*mismatches or undetected similarities limit the quality of the mapping results*" [Lanzenberger et al., 2008]. A multitude of alignment techniques operate with various forms of heterogeneity (e.g., syntactic, terminological, semantic) (cf. Section 2.1); but a *residual risk* remains: the pragmatic and structural heterogeneity (cf. Section 1.2).

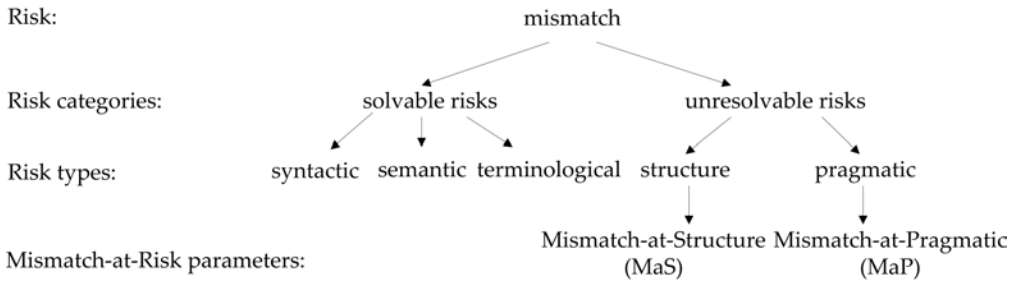



Figure 6.6: Categorization of possible mismatch risks.

Figure 6.6 shows a categorization of risk types, which may cause mismatch when aligning ontologies. In Part B of CoMetO we introduce techniques by which the two kinds of unresolvable heterogeneities are made visible and predictable to users in the form of risk indicators. Firstly, they are made *visible* by the implementation of the iweighting procedures and their output introduced in Part A (cf. Sections 6.3, 6.4); and secondly, they are made *predictable* by the calculations performed by the mismatch-at-risk metrics implemented in this part. It can be assumed that, while experienced users will expect a certain level of these subtle forms of heterogeneity, they have no means to validate their expectations before they actually initiate an alignment process. This leads to *uncertainty* by users about the risk level of a possible mismatch between two sources.

## 6.6 Mismatch-at-Risk Metrics

By introducing risk metrics we propose that an additional aim of alignment support should be to advise users against mismatch caused by heterogeneities, which are currently unresolvable. Therefore, we consider two possible types of mismatch: the first type depends on different cognitive perspectives when describing the same domain of interest, which constitute certain dimensions of context-dependent representations. We call this form *Mismatch-at-Pragmatic ($MaP$)* caused by pragmatic heterogeneity. The second type is resulting from differences of modeling styles or modeling conventions,—the *Mismatch-at-Structure ($MaS$)*, which is caused by structural heterogeneity. The knowledge about the level of these risks prior to starting an alignment would aid users in their decision process in that task. For instance, a low $MaS$ is an evidence for better performing a schema-based technique with a focus on the network structure of the sources. The outcomes of that metrics do not only make the task of schema-based ontology alignment easier to perform, but they also make users better at performing this task, which is a "true meaning of cognitive support" [Falconer et al., 2007, Walenstein, 2002].

### Mismatch-at-risk Metric for Approximating the MaP

For the implementation of the risk metric for approximating the $MaP$ we adapt the technique of *schema fragments* introduced by Rahm et al. [2004] for a user-guided selection process of a *random sample*. In their work they use a fragment-oriented approach to decompose a large

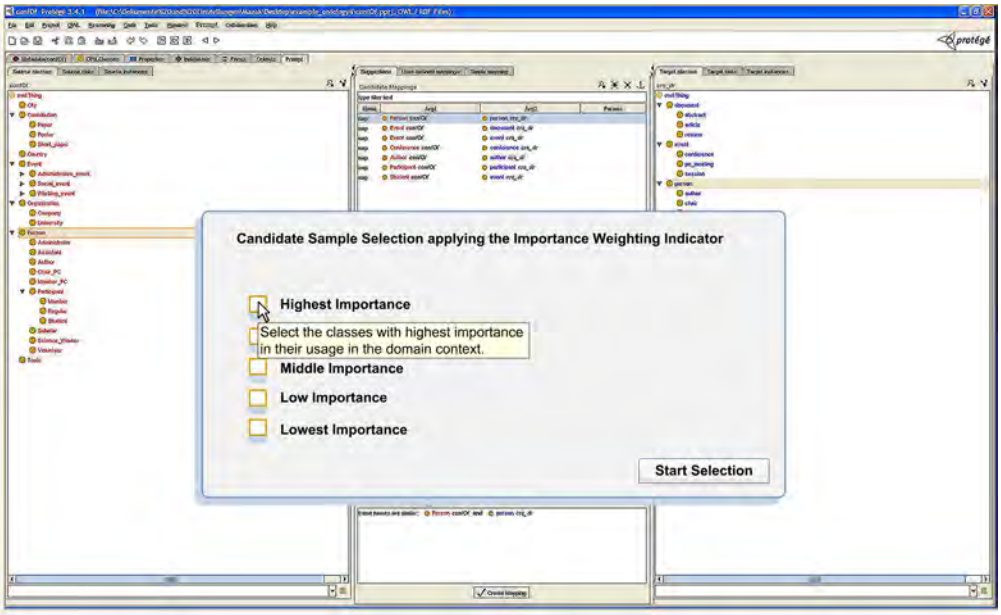



Figure 6.7: Example for selecting a candidate sample using the Anchor-PROMPT Tab Widget plug-in.

matching problem into smaller sub-problems based on a divide-and-conquer strategy, where schema elements become special fragments. In our approach we let the user make a filtering of the $IwI_c$-based ranking lists of two ontologies for identifying a *candidate sample*. That sample is used as fragment and the candidates as *fragment-pairs* (cf. Table 6.5). The mockup, presented in Figure 6.7, shows a simple way for accomplishing a selection of a candidate sample by a point-and-click interaction. For instance, users could select the core concepts of the ranking lists. The interpretation of the input is *external* made by the users that means we leave the final decision to them. Thus, a sample can be classified, according to the schema-matching dimensions described by Shvaiko and Euzenat [2004], as *external resource* in the form of user input. The strategy of a manually conducted candidate selection minimizes the risk of information loss, resulting from possibly poor quality produced by automated methods. The calculations are started after a user has selected a candidate sample of a finite set of classes.

In this section we present the risk metrics by means of the example ontologies $O_A$ and $O_B$. We start with the prediction of a possible pragmatic mismatch on the basis of the iweighted logical statements within each of these ontologies. The direct iweighting procedure was conducted by the participants of our evaluation survey that we present in Chapter 7. We select the core concepts of both ontologies as candidate sample by using the ranking list of Table 6.2a. On the basis of this input we start the calculations. The values we take into account are the accumulated $IwI_c$-based values of Table 7.1 presented in Section 7.2, which represent an arithmetic average of all importance weighting annotations of the survey participants for the classes: *Author*, *Contribution* of $O_A$ and *author*, *article* of $O_B$. These mean values are resulting from the modeling foci of the participants related to each design goal (G1, G2) of the predefined application scenar-

| $IwI_c$ of class | | | | Variation between |
|---|---|---|---|---|
| $O_A$ | | $O_B$ | | $IwI_c$-based values |
| Author | 0.95 | author | 0.93 | 0.00020 $\{x_{11}\}$ |
| Contribution | 0.92 | article | 0.91 | 0.00005 $\{x_{12}\}$ |
| $E_{S1}$ | | | | $\{0.00020,\ 0.00005\}$ |
| Overall unit of risk | | | | 0.000125 |

(a) Measure on the basis of the candidate sample of Sc. 1

| $IwI_c$ of class | | | | Variation between |
|---|---|---|---|---|
| $O_A$ | | $O_B$ | | $IwI_c$-based values |
| Author | 0.08 | author | 0.93 | 0.36 $\{x_{21}\}$ |
| Contribution | 0.11 | article | 0.91 | 0.32 $\{x_{22}\}$ |
| $E_{S2}$ | | | | $\{0.36,\ 0.32\}$ |
| Overall unit of risk | | | | 0.34 |

(b) Measure on the basis of the candidate sample of Sc. 2

Table 6.5: Realizations of $X$ based on Sc. 1 and Sc. 2.

ios; these goals are similar in Scenario 1 (Sc. 1), while dissimilar in Scenario 2 (Sc. 2). The two selected candidate samples, one for each scenario, have the same size as depicted in Table 6.5.

The *random experiment* is performed by a one-to-one comparison of the $IwI_c$-based values of each fragment pair. Thereby, we measure the *variation* between that values by the *variance*. Each fragment pair has size $n = 2$, therefore we use the adjusted variance, without reference to the *measure of location* (i.e., mean value), for computing the variation as representation of pairwise differences [Filzmoser, 2003];

$$s^2 = \frac{(x_1 - y_1)^2}{2}$$

The set of all possible outcomes $\varpi$ of such a measure is denoted as *sample space* $\Omega$ [Dutter, 2002]:

$$\Omega = \{\varpi_1,\ \varpi_2,\ \ldots,\ \varpi_n\}\,;\ \forall\,\{0 \leq \varpi_i \leq 1\}$$

That space is uncountable infinite, because the possible outcomes are in the range of $[0,\ 1]$. The outcome of a single experiment is defined as *elementary event* $\{\varpi\}$ of $\Omega$ [Stahel, 2000]. Certain elementary events can be combined to subsets of $\Omega$. Such subsets are denoted as *events (E)* [Dutter, 2002];

$$\wp(\Omega) = \{E|\ E \subset \Omega\}$$

where $\wp(\Omega)$ is the *event space*, which is the power set (i.e., set of all subsets) of $\Omega$ [Dutter, 2002].

The *spread* or *range of variation* constitutes the *risk indicator* to such a degree as the broader the range is the higher will be the pragmatic heterogeneity risk level. We use this risk indicator as statistically exploitable feature, as an evidence, for a possible pragmatic-based mismatch between the source schemas. The variance is a summable measure, but not a risk measure in a conventional sense, as for instance the standard deviation. It is an indicator which facilitates to quantify uncertainty in that if there is no variation between the pair values (i.e., both $IwI_c$ values are equal) then the variance is zero; but it cannot be less than zero [Stahel, 2000]. The random experiment is a real measuring where the elementary events are real numbers. Therefore, we use the statistical construct of a *random variable*. This construct makes it feasible for us to model the risk of a possible mismatch as a random variable $X$ in order to quantify the uncertainty of such a risk in a numerical value. A random variable is not a variable, it is a function [Fahrmeir et al., 2003];

$$X : \varpi \mapsto x = X(\varpi)$$

by which each elementary event $\{\varpi\}$ of $\Omega$ is assigned to exactly one real number $x$; $x$ is denoted as realization of $X$ [Dutter, 2002].

$$X : f\left[IwI_c(O_A),\ IwI_c(O_B)\right] \rightarrow \mathbb{R};\ x_{ij} \in \mathbb{R} \mid \{0 \leq x_{ij} \leq 1\} \tag{6.3}$$

Table 6.5a presents the measured outcome based on Scenario 1, whereas Table 6.5b presents the outcome of the same sample based on Scenario 2. We index $x_{ij}$ of $X$ in order to relate the realizations to each scenario ($s_i$), and summarize them as subsets ($E_{S1}$, $E_{S2}$). The computations show that the range of variations between the fragment pairs in Scenario 1 is marginal, which is an evidence for a low mismatch risk, whereas that in Scenario 2 is significantly higher.

In a next step we compute the *pooled* or *averaged variance* [Filzmoser, 2003] by aggregating the variation of each fragment pair and take the mean value;

$$S_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2)-2}$$

in order to have an *overall unit of risk* associated with a sample.

The main interest is not the contingency controlled outcome (i.e., overall unit of risk) resulting from the random variation of $IwI_c$-based pair values of a certain candidate sample. The (random) variable we are interested in is a predictor of the probability of an *adverse variation* of the risk factor between the schemas to be aligned; in that the range of variation increases, the probability for a pragmatic-based mismatch grows. Such a parameter, which we denote as $MaP$, functions as an estimator by which the risk level of a possible pragmatic mismatch can be calculated as percentage. For this purpose we have to infer from the realizations of the candidate sample to assumptions regarding to the *population*, which are all indicator-based fragment pairs between the sources. Thereby, $X$ can be used akin to descriptive statistics as a *feature* of a random experiment [Fahrmeir et al., 2003]. In order to approximate the $MaP$ we have to shift the focus by making the variation to a measure of distribution of $X$.

Risks can be evaluated by describing them using an appropriate density, or (probability) distribution function [Franke et al., 2004]. In our approach $X$ is a continuous random variable, since it can take all numerical values of an interval of real numbers. Therefore, the probability distribution of $X$ is given by the *probability density function*. In probability theory the density function $f(t)$ is a function that specifies how significantly the probability of $X$ is concentrated at a certain point $x$ [Stahel, 2000]. The probability that $X$ is not exceeding $x$ is formally defined by the *cumulative distribution function* $F$ which is the integral of the density $f(t)$ [Dutter, 2002]:

$$F(x) = P\left(X \leq x\right) = \int_{-\infty}^{x} f(t)\, d(t)$$

The characteristic parameters, *population mean* $\mu$ and *population variance* $\sigma^2$ of a distribution of $X$ can be estimated by the parameters of the sample, which are the sample mean $\overline{x}$ and the sampling variance $s^2$; or by making assumptions about the family of probability distributions of $X$ and applying the *maximum likelihood* estimate[55] [Filzmoser, 2003]. For instance, by the expected value $E(X) := \mu$ the true mean value, i.e., location of the distribution of realizations

[55] A commonly used method for obtaining an estimate of an unknown parameter of an assumed distribution.

of the population can be described, while the *standard deviation* $\sigma_X := \sqrt{Var(X)}$ describes the deviation from that value. These parameters constitute a measure of the complete set of objects of interest (i.e., all $IwI_c$-based classes of the two sources to be aligned) of a population without really performing it [Stahel, 2000].

In our approach the size of the candidate sample is too small to estimate the characteristic parameters of a distribution of $X$ by the metadata. Therefore, we adapt a "distribution-free" metric from financial statistics,—the *Value at Risk*[56] ($VaR$) metric. "Distribution-free" or "non-parametric" means that the statistical properties of the procedure do not depend on the underlying distribution being sampled, which means that there exists no assumption about the population under investigation [Dutter, 2002].

In financial statistics the $VaR$ metric is a measure of the potential loss of financial positions in a portfolio for a specified period (i.e., holding period) [Eller et al., 2002];

$$VaR = BW \cdot Vola(r_D) \cdot factor_{confidence\ level} \cdot \sqrt{T_D}$$

| Labeling: | |
|---|---|
| $VaR$ | Value-at-Risk according to the rate of a financial instrument |
| $BW$ | actual cash value of that instrument |
| $Vola(r_D)$ | volatility (risk factor) of the rate for a certain time period $D$ |
| $factor_{confidence\ level}$ | scaling factor which constitutes the confidence level |
| $T_D$ | observation period (e.g., 251 days) |

We adapt that part of the $VaR$ metric by which an adverse variation of the underlying risk factor is computable. That random variable is the *volatility* of the rate of a financial instrument (e.g., interest rate) in a certain time period [Eller et al., 2002];

$$\Delta r_D = Vola(r_D) \cdot factor_{confidence\ level}$$

| Labeling: | |
|---|---|
| $\Delta r_D$ | adverse variation of the interest rate risk |
| $Vola(r_D)$ | volatility (risk) of that rate (p.a.) |
| $factor_{confidence\ level}$ | scaling factor which constitutes the confidence level |

Our adaption of this metric is a hybrid-based approach. We combine the historical-based simulation of volatilities, where market price changes over a historical observation period (e.g., 251 days) are used for calculation, with the variance-covariance approach. We denote that metric as *Mismatch-at-Risk* metric. In this section where we cover the pragmatic-based mismatch the outcome of that metric is the *Mismatch-at-Pragmatic* ($MaP$). This parameter is an estimator, such that the probability that the variation gets "unfavorable" because it exceeds this value, is the given value. The $MaP$ metric makes it feasible to compute the probability of such an adverse variation of the risk factor from an expected value as a kind of maximum risk. In financial statistics the volatility is defined as the average deviation of realizations of rate changes from their expected value $E(X) = \mu = 0$; with the assumption that they are *standard normal distributed* [Eller et al., 2002]. The $VaR$ is a *downside* risk measure with a one-sided confidence interval; in that only potential losses are calculated [Franke et al., 2004].

[56] A standard method for measuring market risks as potential loss developed by J.P.Morgan (1996) [Eller et al., 2002].

| $IwI_c$ of class | | | | Variation between |
|---|---|---|---|---|
| $O_A$ | | $O_B$ | | $IwI_c$-based values |
| Author | 0.95 | author | 0.93 | 0.00020 |
| Contribution | 0.92 | article | 0.91 | 0.00005 |
| Overall unit of Sc. 1 | | | | 0.000125 |
| $Var_{IwI}$ based on Sc. 1 | | | | 0.01118 |

(a) Measure on the basis of the candidate sample of Sc. 1

| $IwI_c$ of class | | | | Variation between |
|---|---|---|---|---|
| $O_A$ | | $O_B$ | | $IwI_c$-based values |
| Author | 0.08 | author | 0.93 | 0.36 |
| Contribution | 0.11 | article | 0.91 | 0.32 |
| Overall unit of Sc. 2 | | | | 0.34 |
| $Var_{IwI}$ based on Sc. 2 | | | | 0.58 |

(b) Measure on the basis of the candidate sample of Sc. 2

Table 6.6: Computed heterogeneity factors based on Sc. 1 and Sc. 2.

In our approach we have a random variable, which varies not over time, but from one class to another class. In our adapted form we calculate the $MaP$ on the basis of the computed averaged, or pooled variance of the candidate sample (cf. Table 6.5). We take this overall unit of risk as a kind of "effect size" for the standard deviation, which is equal to the positive square root of the variance of a random variable $\sigma_X := \sqrt{Var(X)}$.

$$Var_{IwI} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2) - 2}} \tag{6.4}$$

We denote the $Var_{IwI}$ as a *heterogeneity risk factor*. A low $Var_{IwI}$ indicates that the $IwI_c$-based values are very close, whereas a high $Var_{IwI}$ indicates that the metadata spread out over a large range, which leads to a high pragmatic-based risk level when aligning the sources. We use that adjusted form of the standard deviation (cf. Equation 6.4) as a statistical measure to estimate how broad the range of variation, which constitutes the risk indicator of the metric, possibly gets. Table 6.6 shows the results of the computations for each scenario. The $Var_{IwI}$ based on Scenario 1 has a very low value, while that based on Scenario 2 indicates a potential high risk level. This heterogeneity risk factor constitutes only an approximation of reality, but the value is sufficiently precise to be used as initial value for estimating the $MaP$ between the sources. It is an actual value by which the level of certainty of a presumption of an unfavorable variation can be approximated.

The $VaR$ metric does not depend on assumptions about the probability distribution of (future) losses; instead it approximates the probability of adverse changes (i.e., possible losses) by *quantile* (i.e., a point with a specified probability $q = 0 < q < 1$) [Franke et al., 2004]. Therefore, this metric is a downside risk measure where only unfavorable variations from an expected value $E(X)$ are considered, which constitute the risk. For instance, by the standard deviation both, positive and negative variations are measured that means "chance and risk". In order to perform the calculations based on this technique we assume that the variations are normal distributed; and convert the random variable $X$ with its (unknown) parameters $\mu$ and $\sigma$ to a standard normal distributed random variable $Z$ with expectation $E(Z) = \mu = 0$ and $\sigma = 1$. For this purpose we use the standardizing transformation [Meintrup and Schäffler, 2005]:

$$Z = \frac{X - \mu}{\sigma}$$

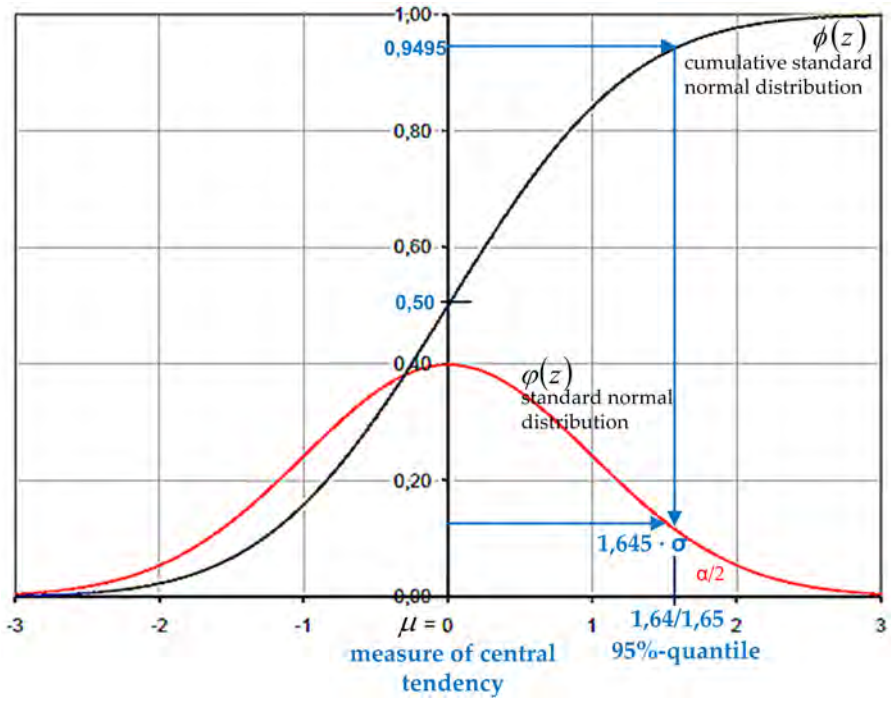



Figure 6.8: Density vs. distribution function of the standard normal distributed random variable.

An expected value close to $0$ means that we expect almost a lowest range of variation between the fragment pairs on average, which implies that the alignment of the sources' schemas bears a minimum risk of pragmatic mismatch. The standardization makes it easy for us to determine a *scaling factor* (i.e., $z - value$) for a certain confidence level (e.g, $95\%$ or $99\%$) as a numarical value by a table, since the distribution function of a normal distributed random variable by which the integral of the density $f(t)$ has to be calculated is not easy to perform. Such a numerical table of values of the standard normal distribution is given in the appendix of every statistical handbook.

Figure 6.8 illustrates the *bell-shaped curve* of the density function $\varphi(z)$ and the *S-shaped curve*, which characterizes the distribution function $\phi(z)$ of the standard normal distributed random variable $Z$. The location parameter $\mu$ constitutes the measure of central tendency. It indicates the point of the distribution function below which $50\%$ of the realizations $z$ of Z lie; thus, $0$ is the $50\%$-quantile [Dutter, 2002]. The *cumulative probability* for a certain confidence level can be easily computed from the positive range of values [Eller et al., 2002]. We marked the cumulative probability with a value of $0.9495$ for the $95\%$-confidence level to demonstrate that the values of $Z$, which are related to this cumulative probability, can be easily determined by the S-shaped curve. According to this the $95\%$-*quantile* lies in the range of $[1.64,\ 1.65]$:

$$P(Z \leq z) = \phi(z)$$
$$P(Z \leq 1.645) = 0.95$$

For arbitrary normal distributed random variables $X$ the quantiles can be determined by the

| $IwI_c$ of class $O_A$ | | $O_B$ | | Variation between $IwI_c$-based values |
|---|---|---|---|---|
| Author | 0.95 | author | 0.93 | 0.00020 |
| Contribution | 0.92 | article | 0.91 | 0.00005 |
| Overall unit of Sc. 1 | | | | 0.000125 |
| $Var_{IwI}$ based on Sc. 1 | | | | 0.01118 |
| $MaP$ by a confidence level of 95% | | | | 2% |

(a) Calculations on the basis of the domain context in Scenario 1

| $IwI_c$ of class $O_A$ | | $O_B$ | | Variation between $IwI_c$-based values |
|---|---|---|---|---|
| Author | 0.08 | author | 0.93 | 0.36 |
| Contribution | 0.11 | article | 0.91 | 0.32 |
| Overall unit of Sc. 2 | | | | 0.34 |
| $Var_{IwI}$ based on Sc. 2 | | | | 0.58 |
| $MaP$ by a confidence level of 95% | | | | 95% |

(b) Calculations on the basis of the domain context in Scenario 2

Table 6.7: Mismatch-at-Pragmatic on the basis of the 95%-quantile.

dispersion parameter $\sigma$ [Eller et al., 2002]. For this purpose the quantiles of the standard normal distributed random variable $Z$ are used as scaling factor with which $\sigma$ is multiplied (e.g., $1.645 \cdot \sigma$) in order to quantify the quantile of $X$, which corresponds to that factor [Eller et al., 2002].

We adapt this procedure for the computations of the $MaP$, as well as the $MaS$ metric. We approximate the $MaP$ similar to the $\Delta r_D$-approach for calculating an adverse variation of an interest rate risk [Eller et al., 2002] as follows:

$$MaP = Var_{IwI} \cdot z - value \tag{6.5}$$

We calculate the Mismatch-at-Pragmatic ($MaP$) as a mismatch rate risk on the basis of the computed heterogeneity factor $Var_{IwI}$, which is the square root of the overall unit of risk $\sigma$, and the scaling factor ($z - value$) of a 95%-quantile, which has a value of 1.645. We multiply the heterogeneity factor with this scaling term in order to approximate the $MaP$ in percentage. The $MaP$ indicates a user the probability of the pragmatic-based heterogeneity getting "unfavorable" when aligning the schemas, because it exceeds this value. This means that with a probability of 95% a mismatch between the sources to be aligned is not exceeding the calculated value, or, that it will be exceeded in 5% of the cases.

We present the final results of the calculations in Table 6.7. We start in our appraisal with the table on the left side (cf. Table 6.7a). The very low variations between the value pairs indicate that the importance of that classes related to their domain context-based usage is very similar. Thus, a user can infer that the modelers' intended meaning on that classes is similar. The heterogeneity factor ($Var_{IwI}$) has a value of $\approx 0.02$, which constitutes a very low risk factor according to a possible pragmatic incompatibility. In contrast, the computed heterogeneity factor based on Scenario 2 (cf. Table 6.7b) has a value of 0.58, which indicates that the $IwI_c$-based values are spread out over a large range. Therefore, a user can infer that the classes' meaning in use is rather dissimilar compared to Scenario 1. On the basis of these heterogeneity factors we calculate the $MaP$ for both scenarios by a 95%-confidence interval. The level determines a scaling factor of 1.645 for further calculations. On the basis of this $z - value$ we approximate a very low $MaP$ in Scenario 1, whereas the predictor for Scenario 2 is highest with a value of 95%. Thus, we can conclude that it would be more favourable to align the schemas of the ontologies in Scenario 1 than those in Scenario 2 in order to minimize a pragmatic mismatch.

### Mismatch-at-risk Metric for Approximating the MaS

For approximating the structure-based mismatch parameter $MaS$ we use the same risk metric as for computing the $MaP$, but with different underlying risk drivers. The *Mismatch-at-Structure* metric is performed on the classes' outdegree by which their absolute frequency as `rdfs:domain` in (logical) statements is inferable in relation to the total number of classes within the ontology. The outcome of that metric provides for users an insight into the modeling context of ontologies.

The risk to be modeled between two schemas is the mismatch caused by differences in the ontologies' structure, which is based on different modeling styles (cf. Sections 1.2, 6.2). For instance, the example ontologies, which both describe the same domain of interest, are different in their structural design: $O_A$ is deeply structured, while $O_B$ has a flat structure as depicted in Figure 3.3 in Section 3.3. A deep structure is an evidence for more `rdfs:subClassOf` relations among classes, whereas a flat structure indicates more binary relations (`owl:ObjectProperty`).

The method for approximating the $MaS$ parameter is in its first calculation step based on the actual outdegree of a class computed by the `setOutdegree`-algorithm (cf. Appendix A.2). The calculated outdegrees of all domain classes are cumulated to a total sum. In a next step this sum is assessed in relation to the number of classes per source; thereby the relative frequency of classes in the role of a domain class is computed in relation to the total number of classes within the schema. This ratio value is calculated as follows:

$$Rf_n(C) = \frac{\sum_{c_i \in C_{Dom}} d^+(c_i)}{|C|},\ C_{Dom} = \{c_i \mid f_n(c_i) > 0\}\,,\ C_{Dom} \subseteq C \tag{6.6}$$

The relative frequency is a statistical approximation by which the probability of a variation can be calculated. The $Rf_n(C)$-ratio is a risk indicator similar to the overall unit of risk computed in the $MaP$ metric. The range of variation between the ratios of two sources, which is the random variable we are interested in this metric, constitutes the risk factor. We compute this factor as a variation of the pairwise difference between the sources' ratios ($x_1$, $y_1$) by the square root of the adjusted variance;

$$Var_{IoI} = \sqrt{\frac{(x_1 - y_1)^2}{2}} \tag{6.7}$$

and denote it as $Var_{IoI}$ following the $Var_{IwI}$.

Table 6.8 presents the result of the $MaS$ metric based on the original authors' modeling conventions when describing the concepts of the domain in $O_A$ and $O_B$. We compute the $MaS$ as a more unfavorable variation of the risk factor with a probability of $5\%$ by the scaling factor of a $95\%$-quantile. The computation is akin to that of the $MaP$-metric (cf. Equation 6.5), which we introduced in detail in the previous section. In our example scenario, the computed $MaS$ parameter gives users an evidence that there exists a $5\%$ probability that the structure-based heterogeneity gets more unfavourable than $79\%$ when aligning the schemas (e.g., when using a graph-based technique).

If we had additionally weighted the taxonomic structure of the ontologies by conducting the indirect importance weighting procedure (cf. Section 6.2), it would be unfeasible to identify

| **class** $O_A$ | $d^+(c_i)$ | **class** $O_B$ | $d^+(c_i)$ |
|---|---|---|---|
| Person | 3 | article | 4 |
| Contribution | 2 | author | 2 |
| Administrative_event | 2 | program | 2 |
| Working_event | 2 | chair | 2 |
| Organization | 2 | abstract | 1 |
| Member_PC | 2 | reviewer | 1 |
| Author | 1 | review | 1 |
| Scholar | 1 | conference | 1 |
| — | — | participant | 1 |
| $\sum_{c_i \in C_{Dom}} d^+(c_i)$ | 15 | | 15 |
| $\lvert C \rvert$ | 38 | | 14 |
| $Rf_n(C)$ | 0.39 | | 1.07 |
| $Var_{IoI}$ | | 0.48 | |
| $MaS$ by a confidence level of 95% | | 79% | |

Table 6.8: Mismatch-at-Structure on the basis of the 95%-quantile.

structural differences between two sources as presented in this section. The consideration of both structures (taxonomic and relational) would result in biased outcomes. The metrics implemented in Part B of align++ are simple yet effective. The statistical methods for approximating a possible structure- and pragmatic-based mismatch between two domain ontologies are well-defined, and as such they provide reliable predictors.

## 6.7 Concluding Remarks

We consider the main components of ontologies, which are: *syntactical features*, *semantic features*, and *pragmatic features* by the methodological part of CoMetO. From the first features we take the underlying graph topology of an ontology (the outdegree of a class and the arcs with their nodes); from the second features we take the type of the model (domain ontology), and the `owl:ObjectProperty` constraints (i.e., `ObjectPropertyDomain`, `ObjectPropertyRange`), and from the pragmatic features we take the expert meta-knowledge of the entities' usage in certain contexts. The outcomes of the method align++ support an "intended meaning negotiation" from ontology engineers to users as proposed by Bouquet et al. [2002] (cf. Section 1.2). By applying the new method users gain context-based evidence from ontology authors in order to get a better understanding of the sources prior to initiating their alignment. For instance, the method could be applied by users in a "pre-alignment phase" to acquire information for candidate selection when searching for potential mappings on

the basis of the indicator-based ranking lists (cf. Section 6.4).

"*A good model depends on the domain of interest, the used ontology language, and the modelers, which reflect the syntactic, semantic, and pragmatic quality*" [Lindland et al., 1994]. For instance, syntactic and semantic quality can be checked by a reasoner (e.g., Pellet[57], KAON[58]), which supports logical formalism. A pragmatic quality checking, which is more usage- or context-dependent is better performed by the engineers' themselves. For this purpose the iweightings in combination with competency questions are a useful aid for pragmatic quality checking (cf. Section 6.1). Bouquet et al. [2004] point out that "*the intended usage has a great impact on alignment, as it can be quite risky to map entities onto each other only because they are semantically related*". We presented two weighting procedures by which modelers are aided to express their intentions regarding the importance of statements and their involved classes in certain contexts.

The motivation of this thesis is to consider the diversity of perspectives in ontology engineering that causes certain heterogeneity risks among the sources when aligning them. align++ is a contribution to improve the quality of schema-based alignment in that we make users aware of the risk level of a possible pragmatic and/or structural mismatch between two sources that both describe the same domain of interest. We presented that the context-based parameters by which a use-conditional form of meaning interpretation is supported, additionally function as risk indicators for that heterogeneities. The first parameter ($IwI_c$) contains information about the classes' usage in the domain context by which a possible pragmatic heterogeneity can be indicated, whereas the second parameter ($IoI_c$) functions as an indicator of a heterogeneity risk resulting from differences in describing concepts (hierarchical vs. network structure). In the *ex post* part of align++ (cf. Section 6.5) we presented that it is feasible to exploit both indicators as statistical features. For this purpose we introduced two mismatch-at-risk metrics where we use constructs of the probability theory and financial statistics. The mismatch parameters, computed on the basis of that metrics, could be an aid for users to select those schema-based alignment techniques that best fit to the approximated values in order to gain better alignment results. We developed a system which makes it feasible to encode (by informal constraints) and decode (by indicators) cognitive semantics in order to provide a shared cognitive environment additionally to a shared physical one. We assume that this approach is a beneficial contribution to improve the reuse potential of newly designed ontologies.

In the following chapter we present the result of our evaluation survey. We discuss that the randomly seeming effects of the computations in Part B are systematical ones by representing the outcomes of the direct iweighting procedures, which were manually conducted by the participants. Additionally, we formulate and test a hypothesis based on that outcomes and the predefined application scenarios.

[57] `http://clarkparsia.com/pellet/` (last accessed July-1-2011)
[58] `http://kaon2.semanticweb.org/` (last accessed July-1-2011)

CHAPTER 7

# Evaluation Survey

In this chapter we underpin our research assumptions, made in the course of the introduction of the CoMetO methodology, by an evaluation survey and a hypothesis testing of Part A of the method align++. We decided to conduct the survey by using a questionnaire in the form that the participants directly can fill in their feedback on the computer. We targeted users and developers with experiences in semantic technologies. We invited 20 persons to participate by e-mail. We defined a time frame of 3 weeks for giving a response. We received a response of 18 persons of the 20 contacted ones; 5 female and 13 male completed the questionnaire, which comprised two sections. We made the questionnaires anonymous before starting the analysis process. 12 of these 18 participants are researchers in Computer Science, while 4 respondents are students in the fields of Computational Intelligence, Software & Information Engineering, and Information & Knowledge Management. Further, 2 respondents are employees in leading positions at a software house. The age of the participants ranges from 25 to 40 years. 12 respondents declared themselves to be well-versed in ontology engineering and alignment, while the others declared themselves as versed; nobody declared herself/himself as unversed in these fields. The respondents were representative for our survey, insofar that they had both academic and industrial background.

## 7.1 Questionnaire Design

We decided to conduct a *closed survey* in order to find representative participants that were able to give a comprehensive feedback. We started with a small demographic block with obligatory entries for age, gender, and profession. Beside a short introduction with explanatory notes we gave evidence about the purpose of the questioning and our expectation regarding the results. The sequence of the questions led from general to particular items. We formulated *direct items* in order to find out facts and wishes as well as *indirect items* for investigating attitudes and perceptions of the participants. We related each question only to one issue; this means that we avoided to ask two things simultaneously. Additionally, we avoided to ask leading questions

as well as double negatives. We asked *close-ended* questions with a fixed set of responses that following a multi-line input box; where the participants could fill in text for giving useful statements, comments, or suggestions to the subject of the asked question. We used Likert[59]-scaled items. Figure 7.1 illustrates the format of the items, which is horizontally structured, rated on a 1-to-5 (strongly disagree - strongly agree) response scale. We decided to take an odd num-

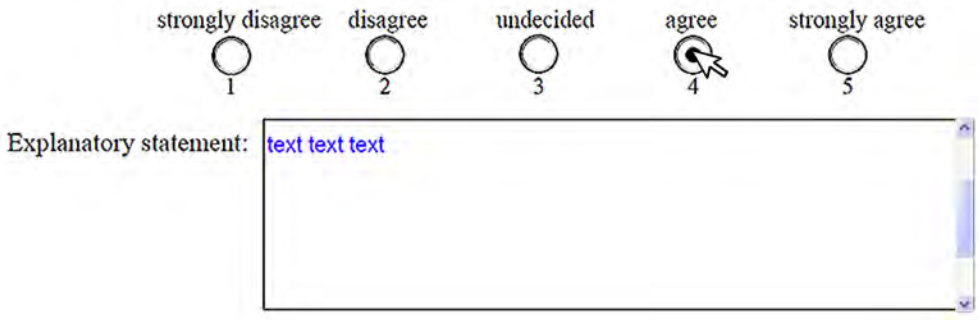



Figure 7.1: Horizontally structured Likert-scaled items.

ber of response categories by providing a neutral alternative item labeled "undecided" in order to prevent omission. The categories complied to the requirements for close-ended questions, which are clearness, completeness, and exclusiveness. In addition, we used a table/matrix (type: single-selection) for the topic "direct iweighting procedure", where multiple issues (e.g., understandability, performance, etc.) were relevant for our analysis, and a dichotomy question (type: yes/no, single-selection) with an additional input box for explanatory statements. Finally, the plausibility of answers was checked by control questions.

## 7.2 Evaluation of the Method align++: First Section

In the course of the first questionnaire section the respondents were asked to weight each `ObjectPropertyAxiom` of the two ontologies ($O_A$, $O_B$) depending on its certain domain/range combinations by annotating an iweighting label. For this purpose we implemented a simple point-and-click user interface for the participants using Excel[60] (cf. Appendix B.1), and we predefined two application scenarios with different design goals ($G_1$, $G_2$) underlying these ontologies:

**Goal of Scenario 1** ($G_1$): both ontologies should be developed to describe the domain concepts *author* and *contribution*. The requirements to fulfill $G_1$, which define $C_{D_1}$, are to support knowledge sharing tasks such as exchanging information on authors (e.g., data of the person, research field) and their submitted contributions (e.g., full paper, short paper, poster, topic) to conferences.

[59] Psychologist Rensis Likert (1903-1981), scaling scheme for measuring personal attitudes.
[60] Microsoft Excel, `http://office.microsoft.com/de-at/excel/` (last accessed July-13-2011).

**Goal of Scenario 2** ($G_2$): the purpose of ontology $O_A$ is to describe the concepts *event* and *organization* of a conference, whereas the purpose of ontology $O_B$ remains the same as in Scenario 1. The requirements to fulfill $G_2$ of $O_A$, which define $C_{D_2}$, are exchanging information on events (e.g., kind of event, temporal order, location) and participating organizations at a conference.

The purpose and expectations of our evaluation survey were:

- to calculate real-valued contextual parameters ($Iw I_c$, $Io I_c$);
- to represent real-valued-based ranking lists as output (cf. Section 6.4, Tables 6.2, 6.4) resulting from that calculated indicators;
- to approximate mismatch-at-risk rates for computing the $MaP$ and $MaS$ caused by the two scenarios (cf. Section 6.5, Tables 6.7, 6.8); and
- to evaluate Part A of align++;

The topic of the questionnaire was derived from these goals. The survey deviates from our approach presented in Chapter 4 insofar that we took two already existing ontologies (`confOf`, `crs_dr`) as basis for the direct iweighting procedure instead of asking the participants to develop new ones from scratch. This means that the ontology design process was only simulated. Firstly, it has shown to be difficult to find representative participants (ca. 20 persons) who are willing to model at least two ontologies based on different design scenario with numerous classes and relations as presented by the example ontologies. Secondly, we compensated this weakness in that we considered the participants in each scenario—together—as a single ontology engineering group by aggregating all of their iweightings and by computing a mean value for the selected samples (cf. Table 7.1) in order to have a valid basis for the computations made in the course of the mismatch-at-risk metrics (cf. Section 6.6). Table 7.1 presents an excerpt of both ontologies described by the classes *Author*/*Contribution* of $O_A$, and *author*/*article* of $O_B$ and their $Iw I_c$-based values. This table shows that all of the 18 respondents weighted the axioms in a nearly equal manner.

Figure 7.2 presents the cumulated distribution for the class *Author* of $O_A$ and *author* of $O_B$ in each scenario. The blue bars show a uniform distribution of iweightings of the class *Author* ($O_A$) in Scenario 1, which implies that all participants weighted this class with an equal importance label; also the class *author* of $O_B$ is approximately uniformly distributed. In Scenario 2 the class *Author* was weighted nearly equal with lowest importance (red bars), only the participants 7, 8, and 11 weighted that class a little bit higher (i.e., low important) in contrast to the other respondents. The plot beside (cf. Figure 7.3) presents a similar tendency of iweighting patterns concerning the classes *Contribution* and *article*. Thus, we can assume that the importance of classes was obviously affected by the participants' modeling focus in $C_{D_1}$ as well as in $C_{D_2}$. The computed mean values of these classes also form the basis for approximating the $MaP$ between the sources described in Section 6.6. Additionally, the mean values of all (domain) classes within the two ontologies are the basis for grouping them in the ranking lists as presented in Table 6.2 (cf. Section 6.4). The predefined design goals were used by the participants as starting point for their iweightings. Usually, such goals are determined by a client or

| Respondent | Ontology $O_A$ Scenario 1 | | Ontology $O_A$ Scenario 2 | | Ontology $O_B$ Both Scenarios | |
|---|---|---|---|---|---|---|
| | *Author* | *Contribution* | *Author* | *Contribution* | *author* | *article* |
| 1 | 0.95 | 0.95 | 0.05 | 0.05 | 0.95 | 0.90 |
| 2 | 0.95 | 0.95 | 0.05 | 0.15 | 0.95 | 0.90 |
| 3 | 0.95 | 0.95 | 0.05 | 0.15 | 0.95 | 0.85 |
| 4 | 0.95 | 0.95 | 0.05 | 0.15 | 0.95 | 0.90 |
| 5 | 0.95 | 0.85 | 0.05 | 0.05 | 0.95 | 0.95 |
| 6 | 0.95 | 0.85 | 0.05 | 0.15 | 0.95 | 0.95 |
| 7 | 0.95 | 0.85 | 0.25 | 0.05 | 0.85 | 0.95 |
| 8 | 0.95 | 0.85 | 0.25 | 0.15 | 0.95 | 0.80 |
| 9 | 0.95 | 0.85 | 0.05 | 0.15 | 0.95 | 0.85 |
| 10 | 0.95 | 0.85 | 0.05 | 0.05 | 0.85 | 0.85 |
| 11 | 0.95 | 0.95 | 0.25 | 0.15 | 0.85 | 0.95 |
| 12 | 0.95 | 0.95 | 0.05 | 0.15 | 0.95 | 0.85 |
| 13 | 0.95 | 0.95 | 0.05 | 0.05 | 0.95 | 0.90 |
| 14 | 0.95 | 0.95 | 0.05 | 0.05 | 0.95 | 0.95 |
| 15 | 0.95 | 0.95 | 0.05 | 0.15 | 0.95 | 0.95 |
| 16 | 0.95 | 0.95 | 0.05 | 0.05 | 0.95 | 0.95 |
| 17 | 0.95 | 0.95 | 0.05 | 0.05 | 0.95 | 0.95 |
| 18 | 0.95 | 0.95 | 0.05 | 0.15 | 0.95 | 0.90 |
| mean value | 0.95 | 0.92 | 0.08 | 0.11 | 0.93 | 0.91 |

Table 7.1: Calculated $IwI_c$-based values for the classes: *Author*/*Contribution* of $O_A$ and *author*/*article* of $O_B$.

a certain application field. Apart from that there existed no direct or indirect influence on our part that enabled us to use the computed values, e.g., to gain a direct comparison between the modeling foci ($MF_{O_A}$, $MF_{O_B}$) in each of the scenarios (cf. Section 7.4).

The plots presented in Figure 7.4 give a visual view of the equalities and differences of the iweightings. In Scenario 1 the participants acted so as to achieve the requirements of $C_{D_1}$ which led to a modeling focus that was mainly on the classes: *Author*, *Contribution* in $O_A$, and *author*, *article* in $O_B$, as well as on these classes' relations to other classes. In $C_{D_2}$ the focus was more on the classes: *Working_event*, *Administrative_event*, and *Organization* in $O_A$. The plot in the top left-hand corner (cf. Figure 7.4a) shows that the participants weighted the classes *Author* ($O_A$) and *author* ($O_B$) based on Scenario 1 in the range of $[0.85,\ 0.95]$, which implies a highest importance of that classes' contextual effect in $C_{D_1}$, whereas based on Scenario 2 they have $IwI_c$-based values in the range of $[0.05,\ 0.25]$, which implies a lowest importance related to $C_{D_1}$. This means that if the modeling focus was mainly on authors and their contributions ($G_1$ of Sc. 1) the relations where these classes are a part in the role of a domain class were weighted highest. Similar can be seen in the example for *Contribution* ($O_A$) and *article* ($O_B$)

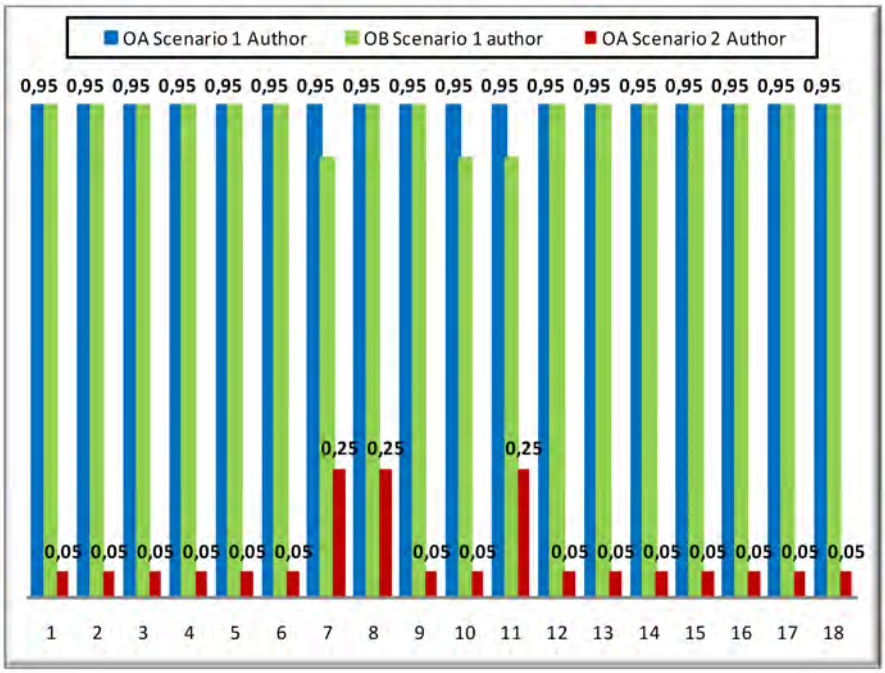



Figure 7.2: Cumulated distribution of $IwI_c$-based values resulting from the participants' iweightings of Author ($O_A$) and author ($O_B$) in each of the scenarios.

(cf. Figure 7.4b). Otherwise, if the focus of $O_A$ was on events and organizations ($C_{D_2}$) the binary relations in which the classes *Author*, *Contribution* participate were weighted lowest, which result in $IwI_c$-based mean values of $0.08$ for *Author* and $0.11$ for *Contribution*. This trend can also be demonstrated by the plot in the bottom left-hand corner (cf. Figure 7.4c), where all participants weighted the class *Administrative_Event* ($O_A$) related to $C_{D_2}$ with an equal iweighting label (i.e., highest importance), which results in an $IwI_c$-based value of $0.95$ for that class. Summing up: There is a very low variation among the $IwI_c$-based values of classes based on Scenario 1, and a very high variation of those values among the same classes in Scenario 2. The latter high variation may lead to mismatch problems caused by pragmatic heterogeneity when aligning the sample sources (cf. Section 6.6).

After the iweighting procedure the participants were asked to answer the survey questions. In the following we present an overview of the ratings and explanatory statements given by the 18 respondents: 89% strongly agree that the modeling focus (MF) on an ontology and its entities depends on a certain perspective ontology engineers have in mind when conceptualizing a domain of interest. They comment that due to semantic relativism, as already known in database engineering, models are always subjective, which cause pragmatic heterogeneity problems in the alignment of these models. 67% strongly agree, and 28% agree that the intended meaning of ontology concepts and their usage mainly depends on the engineers' modeling focus, whereas 5% are undecided. Additionally, they state that the common understanding of engineers which bases on the application of the ontology is important. One of the participants mentions, *"it is*

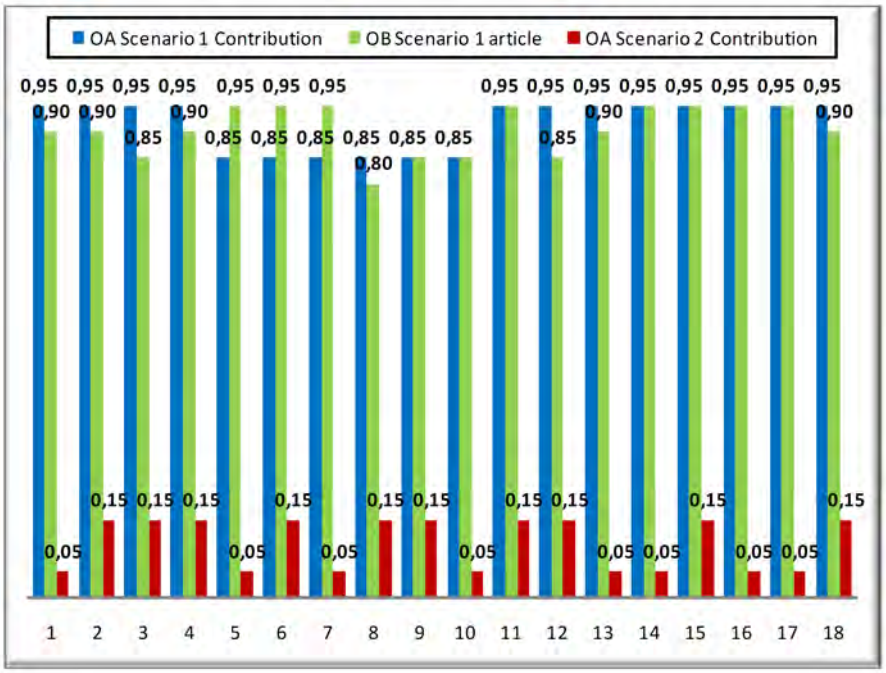
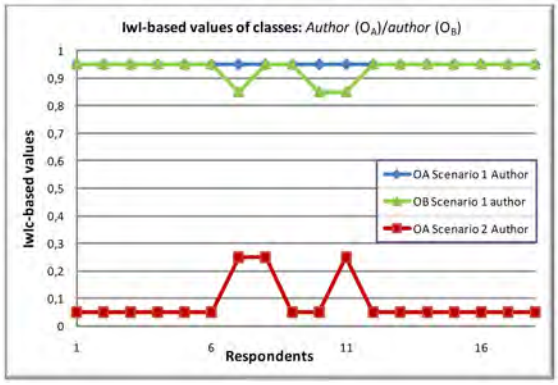



Figure 7.3: Cumulated distribution of $IwI_c$-based values resulting from the participants' iweightings of Contribution ($O_A$) and article ($O_B$) in each of the scenarios.

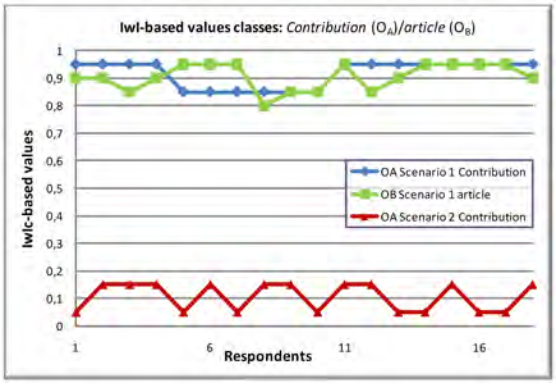
*not possible to model anything without the influence of context-sensitive parameters"*. Another respondent states that *"a concept can be very important in one relation, and unimportant in another depending on the modelers' focus"*. This feedback corresponds to our assumptions and the result of the hypothesis testing (cf. Section 7.4).

Answering the dichotomy question (yes/no) whether there are other components on which the meaning of concepts depends the majority of the respondents replies with "yes". According to the participants these components include *"experiences, culture, stakeholders, background of engineers, skills, environmental parameters, preferences"*. The participants were additionally asked whether they agree that the context-sensitive usage of classes is represented in the logical statements where they are a part. 91% of the participants strongly agree with this assumption. They explain that semantic relations or logical statements are a kind of formalized description of the intended usage of the concepts. The rest argue that also the taxonomic structure, which is commonly used in ontology alignment, should be considered, too. All of the respondents (100%) strongly agree that for instance, the importance weighting degree of the proposition *writes* $\rightarrow$ (*Author*, *Contribution*) would be different if the ontology engineers' modeling focus is on authors rather than on the conference programs. We assume that this "unanimous" answer to the question was influenced by their own experience resulting from the performed iweighting procedure in the first section of the survey.

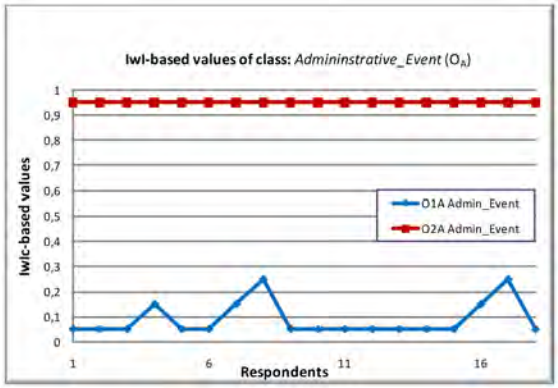
In the align++ approach engineers can choose among five degrees of importance labels in order to add pragmatic-based constraints on propositions:



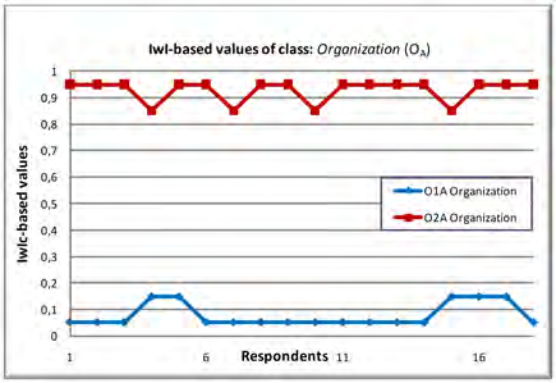

(a) Author ($O_A$) and author ($O_B$)




(b) Contribution ($O_A$) and article ($O_B$)




(c) Administrative_Event ($O_A$)




(d) Organization ($O_A$)

Figure 7.4: Variation among the $IwI_c$-based values of certain classes based on the participants' focus in each scenario.

1. highest importance
2. high importance
3. middle importance
4. low importance
5. lowest importance.

72% of the respondents state that five degrees are enough, 22% consider three as sufficient, and 5% respondent indicate that a finer-grained schema would be better. We hold that five degrees, including a neutral level, are a reasonable compromise. On the issue of whether it is efficient to determine the importance of a particular concept on the binary relations that concept participates in the role of a `rdfs:domain`: 73% strongly argree with this approach, 11% are undecided, and 16% disagree. Those, who disagree point out that only this consideration is not sufficient.

They argue that `is-a` relations, actually meaning `rdfs:subClassOf` relations, should be additionally taken into account when computing a concept's importance.

In the second to last question we pointed to possible heterogeneity problems caused by diversity in ontology modeling based on the engineers' different views of the domain; even though they describe the same domain of interest. The first part of the question was whether it is beneficial that modelers give evidence in the form of importance-based indicators already in ontology development in order to make possible heterogeneities transparent; 22% strongly agree, 33% agree, and 44% are undecided. The second part was better understood in this context. After being shown a brief example that contrasts *Contribution* of $O_A$ and *article* of $O_B$ by their labels and $IwI_c$-based values, 83% strongly agree and 17% agree that an approach where engineers indicate the classes' importance compared to other classes already when developing an ontology would make heterogeneity, such as the pragmatic one, more transparent to users. Additionally, users are made aware of the terminological heterogeneity, which occurs due to variations in names referring to the same concepts. The majority of the respondents point out that it may be useless to align ontologies with different perspectives on their entities. A respondent comments that the calculated indicators "*can immediately give hints, why an alignment would probably fail*". Another participant states that "*they provide an entry point for the alignment process and reduce the probability of wrong perception of the ontology's intended purpose*". All participants strongly agree that the ranking lists are an efficient aid to give users a quick and context-based overview about the core concepts of the sources. They see the benefit in that due to the indicator-based values of classes users can easily detect possible differences in the modelers' foci; as presented by the comparison of the two ranking lists in Table 6.2 (cf. Section 6.4).

Finally, we presented a table/matrix for making an inquiry about the handling of iweighting annotations on statements. We asked for understandability, usability, efficiency, and performance. The summarized results are presented in Table 7.2.

| | unsatisfied | satisfied | very satisfied |
|---|---|---|---|
| Understandability | — | 28% | 72% |
| Usability | — | 22% | 78% |
| Efficiency | — | 22% | 78% |
| Performance | — | 33% | 67% |

Table 7.2: Summary of the inquiry about the handling of the weighting procedure.

## 7.3 Evaluation of the Method align++: Second Section

We started with the second part of the survey after we conveyed the participants' importance weightings in the Odm_ExtensionModelEditor (cf. Section 5.3); on the one hand, to calculate real-valued indicators for Part B of align++ (cf. Section 6.5); and on the other hand, for hypothesis testing. The classes and their calculated $IwI_c$-based values presented in Table 7.1 function as samples for constructing the hypothesis. Hypotheses are often statements about population parameters like expected value $E(X) = \mu$ and variance $\sigma$ [Dolić, 2004].

Firstly, we assume that, based on a different modeling focus, engineers make different design decisions to satisfy the purpose of an ontology, even if they describe the same domain of interest. Secondly, we assume that equal modeling foci on entities of two ontologies imply that the intended meaning between corresponding classes (e.g., *Author*/*author*, *Contribution*/*article*) is similar. If this is not the case, possible heterogeneity risks may occur when aligning the sources, as discussed in Section 6.5.

We hypothesize that in Scenario 1 the modeling foci of the participants are equal on both ontologies:

$$MF_{O_A} = MF_{O_B}$$

whereas in Scenario 2 they are not equal:

$$MF_{O_A} \neq MF_{O_B}$$

We substantiate these assumptions by the results of a parametric hypothesis testing, which we perform in Section 7.4.

## 7.4 Hypothesis Testing

In this section we use the *paired t-test*, which is a parametric test for small paired samples ($n < 30$) for testing the hypothesis, which we made in the aforementioned section. For this purpose we take the classes *Author*, *Contribution* of $O_A$ and *author*, *article* of $O_B$ with their $IwI_c$-based values, presented in Table 7.1, as representative samples. The values are resulting from the modeling focus of each participant when conducting the direct iweighting procedure. Altogether, we have six equally sized samples ($n = 18$); four samples of ontology $O_A$, and two samples of ontology $O_B$ where the predefined design goals were equal for both scenarios. We use R[61] in version 2.9.2 for performing the test for each scenario (`SCENARIO 1`, `SCENARIO 2`). We start the computations by importing the $IwI_c$ values of Table 7.1 as vectors in the R workspace (cf. Figure 7.5). Since, we observed the same group of participants twice under different conditions (Sc. 1, Sc. 2), we perceive these samples as *paired samples*, which means that they are not considered to be independent. The calculation for the *test statistic* $t$ is based not directly on the $IwI_c$-based values, but rather on the differences between these values. The iweightings are a metric feature, therefore we are interested in the differences among that values.

The conditions for performing a paired t-test are as follows:

- randomly selected samples;
- the assumption that the realized values originate from (approximately) normal distributed populations $X_i \sim \mathcal{N}(\mu,\ \sigma^2)$, $Y_i \sim \mathcal{N}(\mu,\ \sigma^2)$; and
- a sample size in the range of: $2 \leq n \leq 30$.

[61] R is a language and environment for statistical computing and available as Free Software, `http://www.r-project.org/` (last accessed June-17-2011).

```
> samples<-cbind(Author_Sc1,author_Sc1,Author_Sc2,Contribution_Sc1,article_Sc1,
Contribution_Sc2)
> samples
      Author_Sc1 author_Sc1 Author_Sc2 Contribution_Sc1 article_Sc1 Contribution_Sc2
 [1,]       0.95       0.95       0.05             0.95        0.90             0.05
 [2,]       0.95       0.95       0.05             0.95        0.90             0.15
 [3,]       0.95       0.95       0.05             0.95        0.85             0.15
 [4,]       0.95       0.95       0.05             0.95        0.90             0.15
 [5,]       0.95       0.95       0.05             0.85        0.95             0.05
 [6,]       0.95       0.95       0.05             0.85        0.95             0.15
 [7,]       0.95       0.85       0.25             0.85        0.95             0.05
 [8,]       0.95       0.95       0.25             0.85        0.80             0.15
 [9,]       0.95       0.95       0.05             0.85        0.85             0.15
[10,]       0.95       0.85       0.05             0.85        0.85             0.05
[11,]       0.95       0.85       0.25             0.95        0.95             0.15
[12,]       0.95       0.95       0.05             0.95        0.85             0.15
[13,]       0.95       0.95       0.05             0.95        0.90             0.05
[14,]       0.95       0.95       0.05             0.95        0.95             0.05
[15,]       0.95       0.95       0.05             0.95        0.95             0.15
[16,]       0.95       0.95       0.05             0.95        0.95             0.05
[17,]       0.95       0.95       0.05             0.95        0.95             0.05
[18,]       0.95       0.95       0.05             0.95        0.90             0.15
```

Figure 7.5: Vector table in the R workspace.

The difference to a normal distribution is that the t-distribution does not depend on $\sigma$, but on $\widehat{s}$ (i.e., the sample standard deviation); and the parameter *df* ($df = n - 1$), which is an `integer` known as the number of degrees of freedom [Dolić, 2004]. Usually, the test statistic follows a *Student's t-distribution* [Dolić, 2004]. The form of that distribution was published in 1908 by Gosset[62], writing under the pen-name "Student" [Dutter, 2002]. With increasing *df* ($df \rightarrow \infty$) the distribution resembles the standard normal distribution.

In a next step, we calculate the differences among the values per class pair (i.e., fragment pair) and participant, and make a brief summary. As example, for the classes *Author* ($O_A$)*/author* ($O_B$) in both scenarios (cf. Figure 7.6). The computed mean in Scenario 1 is very low (0.01667),

```
# SCENARIO 1
> diff_Sc1<-(Author_Sc1-author_Sc1)
> summary(diff_Sc1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.01667 0.00000 0.10000

## SCENARIO 2
> diff_Sc2<-(author_Sc1-Author_Sc2)

>summary(diff_Sc2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.60    0.90    0.90    0.85    0.90    0.90
```

Figure 7.6: Descriptive data analysis of both scenarios.

[62] William Sealy Gosset (1876-1937)

whereas that in Scenario 2 is highest (0.85). The former indicates a low variation among the $IwI_c$-based values, which is an indicator that beside equal labels the "meaning in use" of that classes maybe equal, too. The latter indicates a high variation with obviously dissimilar contextual effects between these classes (cf. Section 6.6).

Before we start with the hypothesis testing in R, we use a null hypothesis ($H_0$) as basis for argumentation of our assumption;

> $H_0$: there is no difference between the modeling focus of $O_A$ and that of $O_B$ on average.

$H_0$ relates to the statement $MF_{O_A} = MF_{O_B}$ being tested, whereas the alternative hypothesis ($H_1$) relates to the statement $MF_{O_A} \neq MF_{O_B}$ to be accepted if $H_0$ is rejected.

> $H_1$: there is a difference on average.

Formally expressed;

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

The alternative hypothesis ($H_1$) is described as two sided ($\mu_D \neq 0$) and the test is *two tailed* [Dutter, 2002]. We test $H_0$ against $H_1$.

The Student's t-distributed test statistic ($t$) is a quantity computed on the basis of the differences between the $IwI_c$-based values of each fragment pair per participant. The difference of each pair constitutes the observation and as that a realization of $D$,

$$D_i = |X_{1i} - X_{2i}|$$

the random variable which we are interested in. The test statistic $t$ is calculated on the basis of [Dolić, 2004];

$$t_{n-1} = \frac{\overline{x}_D - \mu_D}{\widehat{s}_{\overline{x}_D}};\ where\ \widehat{s}_{\overline{x}_D} = \frac{\widehat{s}_D}{\sqrt{n}} \tag{7.1}$$

The parameter $\overline{x}_D$ describes the mean difference of the sample in each of the scenarios, and $\widehat{s}_D$ the samples' standard deviation, being the standard deviation of the differences.

$H_0$ is rejected if:

$$|t| > t_{n-1;1-\frac{\alpha}{2}} \tag{7.2}$$

The level of significance $\alpha = 0.5$, for $\frac{\alpha}{2} = 0.025$. $\alpha$ is the (fixed) probability of a type I error, which occurs when $H_0$ is rejected if it is in fact true, i.e., $H_0$ is wrongly rejected [Dutter, 2002]. The selected significance level of $\frac{\alpha}{2} = 0.025$ leads to a confidence level of 97.5% for the two sided test that results in a *critical value* for the test statistic of [Dutter, 2002];

$$|t| > t_{17;0.975}, \textit{where}$$
$$t_{17;0.975} = 2.110$$

If $T$ has a t-distribution with 17 degrees of freedom then a tabulated value, $t$, is such that

$$P(T < t) = p\%, \text{for } p(97.5\%) = 2.110.$$

This value is a threshold to which the computed (observed) value $t$ of the test statistic is compared to determine whether or not $H_0$ is rejected. The critical value is the boundary of the *rejection region* [Dutter, 2002]. This region is a set of values of a statistic for which $H_0$ is rejected in the testing. Such a critical value for a specified $df$ related to a certain confidence interval (e.g., 95%, 99%) can be easily detected in the appendix of each statistical handbook.

Figure 7.7 presents the result of the paired t-test computed on the basis of the data values of the classes *Author/author* in Scenario 1. The test statistic $t$ is calculated according to Equa-

```
# SCENARIO 1

> t.test(Author_Sc1,author_Sc1,paired=T)

        Paired t-test
data:  Author_Sc1 and author_Sc1
t = 1.8439, df = 17, p-value = 0.0827
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.002403469  0.035736802
sample estimates:
mean of the differences
             0.01666667
```

Figure 7.7: R session of Scenario 1.

tion 7.1, which results in an actual value of $|t| = 1.8439$. This observed value can be interpreted according to Equation 7.2 as follows:

$$|1.8439| < 2.110$$

There is not sufficient evidence against $H_0$ in favour of $H_1$. Thus, $H_0$ cannot be rejected. More intuitively if the actual value (i.e., computed $t$) of the test statistic is close to its expected value the test is deemed to be not significant and the decision is to not reject $H_0$. The test statistic shows that $\alpha < p\ (0 < p < 1)$. The *p-value* constitutes the probability under $H_0$ of observing a value at least as unlikely as the value of the test statistic $t$. If the observed value of the statistic (i.e., the computed value of $t$) is too far from its expected value ($\mu_D$) the test is deemed to be *significant* and the decision is to reject $H_0$ in favour of $H_1$ [Dutter, 2002]. Based on the result

$$p_{0.08279} > \alpha_{0.025}$$

we can conclude that on the level of significance the value of the test statistic is not significantly different from 0, which means that the hypothesis of equal modeling foci in Scenario 1 cannot be rejected.

Figure 7.8 presents the result of the paired t-test based on the data values in Scenario 2. In comparison to the result based on Scenario 1 the computed value of this test statistic is significantly different from 0:

$$p_{2.2e-16} > \alpha_{0.025}$$

```
## SCENARIO 2

> t.test(author_Sc1,Author_Sc2,paired=T)

         Paired t-test
data:  author_Sc1 and Author_Sc2
t = 34.5696, df = 17, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7981236 0.9018764
sample estimates:
mean of the differences
                  0.85
```

Figure 7.8: R session of Scenario 2.

The observed test statistic $|t| = 34.5696$ is significantly greater than the test statistic by a 97.5% confidence level, which means that the actual value is too far from its expected value ($\mu_D$). Therefore, the test is deemed to be significant and the decision is to reject $H_0$ in favour of $H_1$. This result underpins our assumption that the modeling foci in Scenario 2 are not equal. We got the same results when performing the paired t-test on the basis of the classes *Contribution* of $O_A$ and *article* of $O_B$ in each scenario (cf. Appendix C).

## 7.5 Concluding Remarks

Each participant required approximately three hours; firstly, to perform the direct iweighting procedure of all logical statements within the two example ontologies in each of the predefined scenarios (cf. Appendix B.1), and secondly to answer the survey questions and to fill in explanatory statements, subsequent to each of that questions, in the questionnaire (cf. Appendix B). The priority of the survey was given to a high quality in the response. This target and the participants' time effort were the reasons for limiting the study to a small audience; both reasons made it difficult to survey more users to generalize our results to a wider audience.

The response to the questions and the participants' statements revealed that they associated the term "perspective" with a purely subjective focus a modeler has in mind when performing an ontology's development task. They make no distinction between a logical and a cognitive perspective as introduced by Benerecetti et al. [2001] (cf. Section 2.2), and continued in our approach (cf. Section 4.2). Similarly, their understanding of meaning or semantics is a "merged" one. This means that on one hand, they agreed that there exists an intended meaning on concepts based on the modeler's focus at design time; but on the other hand, when aligning two sources, they associated meaning interpretation in a more logic-based sense (as meaning interpretation of language constructs). However, the illustrated results (cf. Figures 7.2, 7.3, 7.4) of the directly, manually conducted iweighting procedures as well as the obtained values (cf. Table 7.1) show that there exists a difference in the classes' meaning in use. This outcome validates the difficulty that even if two classes have the same label, for instance;

$$O_A : \textit{Author} \equiv O_B : \textit{author}$$

a user cannot infer that these classes' meaning in use is similar, too, which leads to pragmatic heterogeneity. We showed that such a problem is mainly caused by the modeling focus in certain contexts ($C_{D_1}$, $C_{D_2}$). The differences in the participants' iweightings encouraged us that users need access to such specific knowledge, which otherwise get no attention, since there exists no entry by model-theory (cf. Section 4.3). For this reason we implemented a new design methodology (cf. Chapter 4), a metamodel (cf. Chapter 5), as well as a method (cf. Chapter 6), which—together—support a use-conditional form of meaning consideration (at design time) as well as an interpretation (when aligning the sources).

The participants agreed that there are other components on that the intended meaning on concepts depends: skills, education, experience, preferences, culture, etc. Their statements coincide with the results of the pilot study conducted by Smart and Engelbrecht [2008] (cf. Section 1.2), as well as other research works (e.g., Bontas [2005], Chalupsky [2000], Falconer and Storey [2007], Klein [2001]) in the field of "cognitive support in ontology alignment". For us it was an important advice that we have to consider more than the modeler's cognitive perspective related to the domain context. Therefore, we revealed in CoMetO to such characteristics of engineers not by an additional perspective, but rather by introducing an additional context,—the *modeling context*.

CHAPTER 8

# Conclusions and Future Work

In this chapter the main contributions of this thesis are outlined and a number of directions for future work are discussed.

## 8.1 Summary of Contributions

The keystones of this thesis are firstly motivated by Saxe's story about the "*blind men and the elephant*" [Saxe, 1887], which we understood as a metaphor to research the variety of perspectives when developing an ontology. The ideas that we introduced are the ones that we considered as useful based on the detailed literature research, our own experience in ontology engineering and alignment, as well as on the evaluation survey. We hold that the implementation of our approach could be best tackled from a multidisciplinary point of view by using techniques from different disciplines: pragmatics, relevance theory, graph theory, financial and inferential statistics, and probability theory. We set to work on the integration of these fields by presenting an approach consisting of three parts:

1. theoretical part,
2. conceptual part,
3. methodological part.

To the best of our knowledge, such an approach has never been studied before and distinguishes our procedural method from the others presented in the state of the art chapter.

In the first part, we introduced CoMetO—a cognitive design methodology for enhancing the alignment potential of ontologies (cf. Section 4.2). We focused on the socio-technical component in ontology engineering (cf. Section 3.1) when we presented a theoretical concept to foster an *evidence-based communication* from engineers to users. We emphasized a better integrated user support in ontology alignment that already starts when developing ontologies (cf. Section 4.3). Such an early consideration of "alignment support" is crucial in our approach as

well as the consideration of the ontology, the modeler's view of the domain, and contexts in a single environment. That differentiates our approach from those works that we outlined in Section 2.2. The theoretical part of CoMetO is mainly influenced by the relevance-based inferential model of verbal communication (cf. Section 4.4). Our idea was to adapt this model in order to supplement the ontology's relational structure with (context-based) *cognitive semantics* to provide—in combination with model-based semantics—a "complete package" for meaning interpretation as input in the alignment process. For this purpose we introduced the concept of a *modeling focus* by which the modeler's cognitive perspective in certain contexts can be represented (cf. Section 4.5). The implementation of this representation formalism makes it feasible for modelers to give information of the entities' *relevance in certain contexts* to users. By the work in the theoretical part of CoMetO we fulfilled our first objective (cf. Section 1.4):

> "*to introduce a representation formalism for the modeler's cognitive perspective in order to make it visible to users when aligning ontologies.*"

In the second part we implemented the theoretical part by a metamodel in that we extended the schema definition language OWL DL by using the Eclipse Modeling Framework (cf. Section 5.3). In the CoMetO metamodel we considered the modeling focus as relation between the logical form and the engineer's cognitive perspective on schema level entities. We used that implementation to join the logical and cognitive perspectives. The access to that *informal expert meta-knowledge* is given by the ontology authors themselves. We hold that nobody can express this knowledge better than those parties who are involved in the design process. For this purpose we implemented a metamodel (cf. Section 5.1) by which such an expression is facilitated as *cognitive constraints*:

- *pragmatic-based constraints*
- *structure-based constraints*

in the form of *importance weighting annotations* (e.g., weighting labels and computed values). By doing so the modeler determines the contextual effect of logical statements and their participating classes in the domain description. In the second part of our approach we presented a metamodel in order

> "*to introduce a method by which the relevance of ontology entities can be evaluated based on their usage in certain contexts*"

which constitutes our second objective.

In the third part, the methodological part of CoMetO, we presented "*align++*"—a method by which a *use-conditional (evidence-based) inference* is provided for users (cf. Section 6.1). Such an inference is facilitated by automatically computed *contextual parameters* that are based on the outcome of two weighting procedures by which the cognitive constraints can be performed (cf. Section 6.3). We introduced two parameters that both indicate information about the (domain) classes' usage in certain contexts (domain and modeling context):

- *Importance Weighting Indicator* ($IwI_c$)
- *Importance Outdegree Indicator* ($IoI_c$)

Additionally, this encoded meta-information can be used to make potential heterogeneities that are currently unresolvable (pragmatic and structural heterogeneity) visible to users. By introducing these indicators we fulfilled another objective:

> "*to generate additional indicator features for classes by which they can be ranked in lists in order to make their originally intended importance visible to users.*"

These ranking lists are an aid to give users a quick and context-based overview of potential mapping candidates (cf. Section 6.4). The *ex post* part of align++ consists of two *mismatch-at-risk metrics*, which we adapted from a risk metric of financial statistics and concepts of inferential statistics. The inputs of those metrics are the indicator-based metadata ($IwI_c$- and $IoI_c$-based values) of Part A. The computed *mismatch-at-risk parameters*:

- *Mismatch-at-Pragmatic* ($MaP$)
- *Mismatch-at-Structure* ($MaS$)

are predictors of mismatch, caused by pragmatic- and structural heterogeneity, which constitutes the final objective in this thesis:

> "*to provide predictors of potential structure- and pragmatic-based mismatch to users prior to starting an alignment process.*"

We presented the computation and outputs of the mismatch-at-risk metrics on the basis of the importance-weighted relational structures of the example ontologies `confOf` and `crs_dr` (cf. Section 6.5).

We concluded our work by a survey where the participants:

1. manually conducted a direct importance weighting procedure of all propositions within the example ontologies based on predefined design goals (cf. Section 7.2);
2. responded a questionnaire and filled in explanatory statements to evaluate the approach of align++; and
3. where we underpinned parts of our research assumptions by hypothesis testing based on the outcome of the participants iweighting annotations and the predefined application scenarios (cf. Section 7.4).

The result of this survey confirms our literature-based analysis and our assumptions made in Chapter 4. The illustrated results (cf. Figures 7.2, 7.3, 7.4) of the directly, manually conducted iweighting procedures as well as the obtained values (cf. Table 7.1) show that there exists a difference in the importance of the classes' meaning in use. The outcome validates the difficulty that even if two classes have the same label (e.g., *Author/author*), a user cannot infer that these classes' usage in context is similar, too. Our aim was to show that such a problem is

mainly caused by the modeling focus of the participants, even though the ontologies describe the same domain of interest (a software tool for conference organization support). Therefore, we hypothesized that in Scenario 1 (where the design goals were the same) the modeling foci of the participants were equal on average, whereas in Scenario 2 (where the design goals were different and so were the domain contexts) they were not equal. The performed paired t-tests significantly confirmed this hypothesis (cf. Section 7.4).

## 8.2 Future Work

In this section we outline directions of future research work related to the results presented in this thesis.

**OWL constructs:** It would be beneficial to extend the approach by including `owl:DatatypeProperty` constructs in an analogous way, since they are not covered by typical schema-based alignment methods. Additionally, other OWL constructs should be considered like class relations (e.g., disjoint sets), cardinality restrictions, as well as the new features (e.g., `ObjectPropertyChain`) introduced in OWL2 [W3C, 2009].

**Evaluation survey:** The weakness of our survey is that the ontologies were not developed from scratch as proposed in the CoMetO design methodology. It would be fruitful to initiate a survey where a group of participants develop at least two ontologies with a representative number of classes and relations among them and another group of participants, who are not part of the original modelers, should be observed when using CoMetO in ontology maintenance. For instance, to get useful hints how they are aided by the outcome of align++ in their understanding of the domain ontologies when they extend or modify that sources.

**Quantitative examination:** *Precision* and *recall* are used to evaluate the quality of mapping candidates [Ehrig, 2007, Euzenat and Shvaiko, 2007]. Ehrig and Sure [2004] comment that "*some mappings are simply not identifiable, not even by humans*". They point out that this is the reason why some results only reach an unsatisfying level of recall. Therefore, it would be beneficial to investigate if the recall rises when align++ is implemented.

**Ontology visualization:** The participants of an evaluation survey conducted by Falconer [2007] pass criticism on (candidate) lists as provided e.g., by PROMPT [Noy and Musen, 2003]. They would find it more useful to navigate through ontology trees instead of reading lists. Thus, another prospect is to integrate the methodological part of CoMetO in a visualizing tool, e.g., AlViz [Lanzenberger and Sampson, 2006] in order to improve the visualization of candidate-heavy regions, which would aid users in their understanding of the sources by browsing such trees.

**Semantic Web:** It would be especially interesting to analyze the usage of iweighted domain ontologies as input to improve the ranking of search results of large-scale Semantic Web search engines.

APPENDIX A

# Source code CoMetO Metamodel

## A.1 getLocalWeight()

```
1  public float getLocalWeight() {
2    float Sum=0.0f;
3    int uOP=0;
4
5    // all used object properties of that class
6    for (Property o: getUsed_Property()) {
7      // Iweights of that properties
8      for (Weighting w: ((OWLObjectProperty)o).getIweight()) {
9        // only that Iweights where the actual class is a domain class
10       if (w.getDomain() == this) {
11          uOP++;
12          switch (w.getValue()) {
13            case LOWEST:
14              Sum += 0.05f;
15              break;
16            case LOW:
17              Sum += 0.25f;
18              break;
19            case MIDDLE:
20              Sum += 0.5f;
21              break;
22            case HIGH:
23              Sum += 0.75f;
24              break;
25            case HIGHEST:
26              Sum += 0.95f;
27              break;
28          }
29       }
```

```
    }
  }

  System.out.println(Sum + " " + uOP);
  return Sum/uOP;
}
```

## A.2 setOutdegree()

```
public void setOutdegree(int newOutdegree) {
  int oldOutdegree = outdegree;
  outdegree = 0;

  Iterator<Property> propertyIterator =
    this.getUsed_Property().iterator();
  while(propertyIterator.hasNext()) {
    Property property = propertyIterator.next();

    if(property instanceof OWLObjectProperty) {
      outdegree++;
    }
  }

  if (eNotificationRequired())
    eNotify(new ENotificationImpl(this, Notification.SET,
      Odm_extensionPackage.OWL_CLASS__OUTDEGREE, oldOutdegree,
      outdegree));
}
```

## A.3 setRatio()

```
public void setRatio(float newRatio) {
  float oldRatio_O = ratio;
  ratio = newRatio;

  try {
    OWLOntology ontology = (OWLOntology) this.eContainer();

    if(ontology.getMaxoutdegree() > 0) {
      ratio = ((Float.parseFloat(this.getOutdegree() + "") /
        ontology.getMaxoutdegree()));
    } else {
      ratio = -1;
    }
  } catch (Exception e) {
    ratio = 0;
```

```
  }

  if (eNotificationRequired())
    eNotify(new ENotificationImpl(this, Notification.SET,
      Odm_extensionPackage.OWL_CLASS__RATIO, oldRatio_O, ratio));
}
```

## A.4 RankIwI()

```
public void setRankIwI() {
  Iterator<OWLClass> classIterator = this.getOWLClasses().iterator();

  ArrayList<String> Lowest_Importance = new ArrayList<String>();
  ArrayList<String> Low_Importance = new ArrayList<String>();
  ArrayList<String> Middle_Importance = new ArrayList<String>();
  ArrayList<String> High_Importance = new ArrayList<String>();
  ArrayList<String> Highest_Importance = new ArrayList<String>();

  while(classIterator.hasNext()) {
    OWLClass owlClass = classIterator.next();

    if(owlClass.getLocalWeight() <= (0.15)) {
      Lowest_Importance.add(owlClass.getName());
    }
    else if(owlClass.getLocalWeight() > (0.15) &&
      owlClass.getLocalWeight() <= (0.25)) {
      Low_Importance.add(owlClass.getName());
    }
    else if(owlClass.getLocalWeight() > (0.25) &&
      owlClass.getLocalWeight() <= (0.50)) {
      Middle_Importance.add(owlClass.getName());
    }
    else if(owlClass.getLocalWeight() > (0.50) &&
      owlClass.getLocalWeight() <= (0.75)) {
      High_Importance.add(owlClass.getName());
    }
    else if(owlClass.getLocalWeight() > (0.75)) {
      Highest_Importance.add(owlClass.getName());
    }
  }
}
```

## A.5 RankIoI()

```
public void setRankIoI() {
  Iterator<OWLClass> classIterator = this.getOWLClasses().iterator();
```

```
3
4     ArrayList<String> Low_Importance = new ArrayList<String>();
5     ArrayList<String> Middle_Importance = new ArrayList<String>();
6     ArrayList<String> High_Importance = new ArrayList<String>();
7
8     while(classIterator.hasNext()) {
9       OWLClass owlClass = classIterator.next();
10
11      if(owlClass.getRatio() > (0.01) &&
12        owlClass.getRatio() <= (0.40)) {
13        Low_Relevance.add(owlClass.getName());
14      }
15      else if(owlClass.getRatio() > (0.40) &&
16        owlClass.getRatio() <= (0.70)) {
17        Middle_Relevance.add(owlClass.getName());
18      }
19      else if(owlClass.getRatio() > (0.70)) {
20        High_Relevance.add(owlClass.getName());
21      }
22     }
23  }
```

APPENDIX B

# Survey Questionnaire

Annex to Chapter 6

# Evaluation of the method align++



Attachment: example ontologies to manually conduct the importance weighting annotation procedure
*confOf* ontology (A)
*crs_dr* ontology (B)

## PERSONAL DATA

| | |
|---|---|
| **Date** | |
| **Gender** | ☐ female<br>☐ male |
| **Age** | |
| **Field of study** | ☐ student in |
| **Field of research** | ☐ researcher in |
| **University/faculty** | |
| **Field of business activity** | ☐ entrepreneur in<br>☐ employee in<br>☐ other in |

## Short introduction to the method align++

*align++* is a semi-automatic method enhancing the cognitive support[1] for users in ontology alignment. The name align++ results from the two steps in which the method is divided; an ex ante and an ex post step. The method is a hybrid-based approach exploiting the advantages of structure- and element-level techniques[2]. More precisely, the align++ method is a combination of graph- and model-based techniques, and also lexical methods which aligns the concepts as lists where the order of the concepts is not critical. Therefore, we classify align++ as an element-level semantic alignment method[3].

First step of align++: Ontology alignment methods analyze mainly two factors; entity labels and relations among entities. We propose to consider a third factor, the *modeling focus* of ontology engineers. This focus conveys on the one hand, the importance of ontology concepts which derives from the level of their information significance in the modeling context, and on the other hand their importance for structure-based alignment techniques.

The modeling focus on a particular *concept c* can be observed and measured by two indicators: the *importance weighted relation indicator (IwI$_c$)*, and the *importance outdegree indicator (IoI$_c$)*. The *IwI$_c$* of a concept results from the weighted semantics of relations depending on their domain/range combinations (axioms) this concept participates. The weighting annotation of each logical statement is explicitly asserted by the ontology author during the modeling process. They can distinguish between five degrees of iweighting labels. The measuring procedure is a manually conducted weighting function based on a case differentiation. We think that users prefer to assign importance labels instead of numerical values.

| Importance Weighting Label | Description |
|---|---|
| Highest Importance | The logical statement has a highest significance in its meaning in the modeling focus. |
| High Importance | The logical statement has a high significance in its meaning in the modeling focus. |
| Middle Importance | The logical statement has a medium significance in its meaning in the modeling focus. |
| Low Importance | The logical statement has a low significance in its meaning in the modeling focus. |
| Lowest Importance | The logical statement has a lowest signfic-ance in its meaning in the modeling focus. |

Table 1: Iweighting importance degrees and their descriptions.

For instance: We use OWL as vocabulary to describe domains of interest. In the example ontology the modeling focus is on professors and their publications.

[1] S. M. Falconer, N. F. Noy and M.-A. Storey, “Towards understanding the needs of cognitive support for ontology mapping”, OM-2006, Georgia, USA.
[2] J. Euzenat and P. Shvaiko, “Ontology Matching”, Springer-Verlag Berlin Heidelberg, 2007, Fig. 3.1, p. 65.
[3] F. Giunchiglia, P. Shvaiko, “Sematnic Matching“, Technical Report DIT-03-013, 2003, http://eprints.biblio.unitn.it/archive/00000381/01/013.pdf, online checked 08.01.2010.

../4

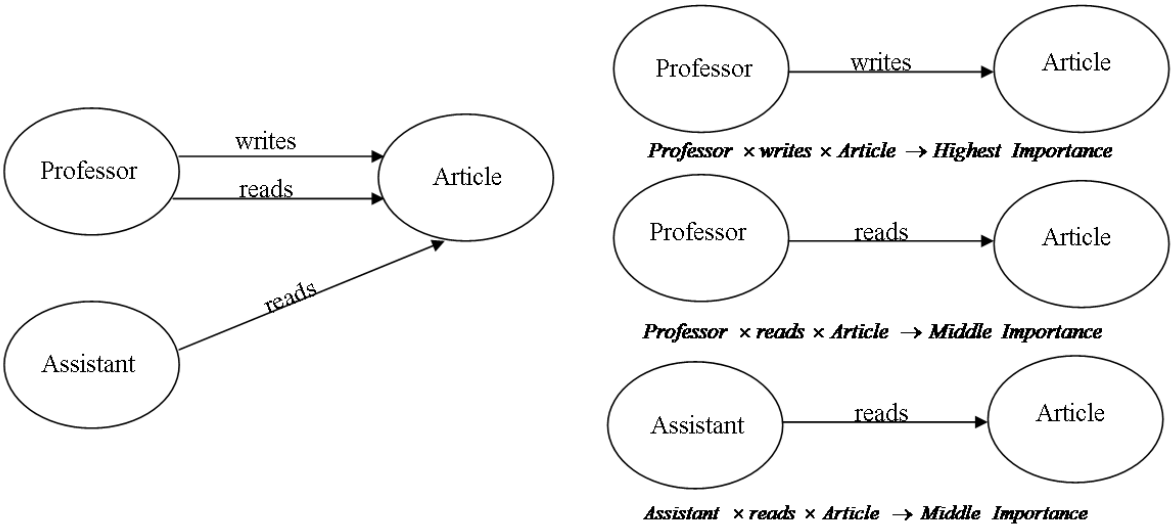



Figure 1: Manually conducted importance weighting measuring procedure by the ontology author.

The logical statement *Professor→writes→Article*, is highly important, while the fact that they *read Article* has not the same importance. This semantic relation has a middle importance, and *Assistant→reads→Article* has also a medium weight in its context-sensitive meaning.

In the next step the method converts each annotated importance weighting degree label to its numerical counterpart in the range of [0.05, 0.95] for further computation.

| **Logical Statement** | **iweighting Degree Lable** | **Numerical Value** |
|---|---|---|
| *Professor→writes→Article* | Highest Importance | 0.95 |
| *Professor→reads→Article* | Middle Importance | 0.50 |
| *Assistant→reads→Article* | Middle Importance | 0.50 |

Table 2: Weighted semantics of the relations among the concepts *Professor*, *Assistant*, and *Article*.

After this step, the algorithm calculates for each concept, in the role of a domain class, an $IwI_c$-based measured mean value.

$$IwI_c = \frac{1}{|OP(x)|} \sum_{i \in OP(x)}^{|OP(x)|} iw_{OPi}^{(x,y)} : \qquad IwI_{Professor} = 0.73 \text{ and } IwI_{Assistant} = 0.50$$

The $IoI_c$-based value results from the number of outgoing relations (outdegree) of a class in proportion to the particular class with the most outgoing relations (highest outdegree) in the ontology.

$$IoI_c = \frac{|OP(x)|}{\max|OP(y)|} : \qquad IoI_{Professor} = 1 \text{ and } IoI_{Assistant} = 0.50$$

In the example ontology the class *Professor* has the highest outdegree with two outgoing relations to another concept.

These indicators are two modes for ranking ontology concepts. Therefore, the output of align++ in its first step are ranked lists of concepts from each source ontology. In these lists the concepts are grouped by their mean value of the importance weighted relation indicator and by the value of their importance outdegree indicator. The $IwI_c$-based ranking lists support

users to detect the core concepts of each source ontology. Additionally, the $IoI_c$-based ranking lists help users to determine efficient candidates for becoming initial points for structure-based alignment methods. The lists are an aid to support users in getting a quick overview of the source ontologies, and an idea about the modeling focus on their concepts.

## PART A

The first part of this survey evaluates the individual *importance weighting (iweighting) annotation process* during the development of an ontology. This iweighting process is the crucial part of the method align++ in its first step. The survey of Part A should detect the core components which impact the setting of *importance weighting degree labels* in this measuring procedure. Additionally, the understandability and usability of the method align++ in its first step should be evaluated.

Please, answer the following questions of Part A:

1 What is your background knowledge in ontology engineering?

| | |
|---|---|
| Well-versed | ☐ |
| Versed | ☐ |
| Unversed | ☐ |

2 You find two example ontologies, *confOf* (ontology A) and *crs_dr* (ontology B), with a predefined modeling focus, and a tool for conference organization support as domain of interest. Please, assign an importance weighting degree to each object property with its certain domain/range combinations depending on the particular (predefined) modeling focus by a simple point and click ⊙ interaction in each case.

| | | |
|---|---|---|
| a) | same modeling focus on both ontologies: | *authors* and *papers* |
| b) | different modeling focus on the ontologies: | ontology A: *events* and *organization*<br>ontology B: *author* and *papers* |

3 Do you agree with the assumption that the modeling focus of an ontology and its entities depends on a certain perspective that ontology engineers have in mind when conceptualizing a domain of interest?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

4 Do you agree that a modeling focus on concepts depends on their context-sensitive usage within an ontology?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

../7

4.1 Do you agree that the meaning of ontology concepts mainly depends on the modeling focus ontology engineers had in mind when conceptualizing an ontology of a certain domain?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

4.2 Do you think that there are other components on which the meaning of concepts depends?

Yes ☐
No ☐

Explanatory statement: .....

.....

.....

5 Do you agree that the logical statements (semantic relations) among concepts are an indicator for their context-sensitive usage?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

6 Do you agree that each importance weighting label only depends on the four components of the quadruple: (modeling focus, owl:ObjectProperty, rdfs:domain, rdfs:range) by applying OWL as vocabulary used to describe domains of interest?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

7 Do you agree that the importance weighting degree of the example relation *author → writes → contribution* would be different if the modeling focus of the ontology engineers was on the *authors* rather than on the *conference program*?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

8 Do you agree that five importance weighting levels: *Highest Importance, High Importance, Middle Importance, Low Importance,* and *Lowest Importance* are sufficient for weighting the logical statements among ontology concepts?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

9 Do you agree that the calculation of an $IwI_c$-based measured value for an ontology concept based on the mean of all semantic relations this concept participates in the role of a domain class (rdfs:domain) to other concepts is efficient to determine the importance of the particular concept in the modeling focus?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

10 Ontology creators have different interests relating to the development of ontologies. Frequently, ontologies based on the same domain of interest are similar but also have many differences which are known as heterogeneity[4]. The reason behind heterogeneity is rooted in diversity in ontology modeling based on different views creators have on a domain. Ontology mismatch in the alignment process is the consequence of this heterogeneity.

[4] J. Euzenat and P. Shvaiko, "Ontology Matching", Springer-Verlag Berlin Heidelberg, 2007, pp. 40-44.

10.1 Do you agree that a method which starts with the measurements of its indicators already during the ontology development process makes heterogeneity more transparent for the (end)user in the alignment process?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

10.2 The concepts *contribution* (ontology A) and *article* (ontology B) are an example of *terminological heterogeneity* between two possible candidates. This kind of heterogeneity occurs due to variations in names referring to the same entities. In this example we assume that the modeling focus is the same in both ontologies. The engineers of the ontologies weight all logical statements with *High* or *Highest Importance* where the concepts *contribution* (ontology A) and *article* (ontology B) participate. Therefore, both concepts have a highest calculated $IwI_c$-based value in the range of [0.75, 1]. The method align++ suggests these two concepts as efficient candidates as a result of their $IwI_c$-based values (*contribution*=0.81, *article*=0.89). Do you agree that these values make heterogeneity problems more easily manageable for (end)users?

| strongly disagree | disagree | undecided | agree | strongly agree |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ |
| 1 | 2 | 3 | 4 | 5 |

Explanatory statement: .....

.....

.....

11 How satisfied are you with the following items according to the importance weighting degree annotation, as first step, in the method align++?

| | unsatisfied | satisfied | very satisfied |
|---|---|---|---|
| Understandability | ☐ | ☐ | ☐ |
| Usability | ☐ | ☐ | ☐ |
| Efficiency | ☐ | ☐ | ☐ |
| Performance | ☐ | ☐ | ☐ |

Comments, suggestions: .....

.....

.....

**THANK YOU FOR YOUR CONTRIBUTION!**

## B.1 Example Ontologies: confOf and crs_dr

Evaluation of the method align++

**Ontology A** **confOf**

**Modeling focus on:** *Authors and their Contributions*

**domain of interest:** tool for conference organization support

**owl:Classes** 38

**owl:ObjectProperties** 13

**iweigthing Importance Degrees** related to the predefined modeling focus

| ID | IS-A Taxonomie | inverse OP | Domain | ObjectProperty | Range | Lowest Importance | Low Importance | Middle Importance | High Importance | Highest Importance |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | super-class | | **City** | | | | | | | |
| 2 | super-class | | **Contribution** | dealsWith | **Topic** | | | | | |
| 3 | | inverse to 34 | **Contribution** | writtenBy | **Author** | | | | | |
| 4 | sub-class | | **Paper** | | | | | | | |
| 5 | sub-class | | **Poster** | | | | | | | |
| 6 | sub-class | | **Short_paper** | | | | | | | |
| 7 | super-class | | **Country** | | | | | | | |
| 8 | super-class | | **Event** | | | | | | | |
| 9 | sub-class | | **Administrative_event** | follows | **Administrative_event** | | | | | |
| 10 | | | **Administrative_event** | parallel_with | **Administrative_event** | | | | | |
| 11 | sub-sub-class | | **Camera_Ready_event** | | | | | | | |
| 12 | sub-sub-class | | **Registration_of_participants_event** | | | | | | | |
| 13 | sub-sub-class | | **Reviewing_event** | | | | | | | |
| 14 | sub-sub-class | | **Reviewing_results_event** | | | | | | | |
| 15 | sub-sub-class | | **Submission_event** | | | | | | | |
| 16 | sub-class | | **Social_event** | | | | | | | |
| 17 | sub-sub-class | | **Banquet** | | | | | | | |
| 18 | sub-sub-class | | **Reception** | | | | | | | |
| 19 | sub-sub-class | | **Trip** | | | | | | | |
| 20 | sub-class | | **Working_event** | hasAdminEvent | **Administrative_event** | | | | | |
| 21 | | | **Working_event** | hasTopic | **Topic** | | | | | |
| 22 | sub-sub-class | | **Conference** | | | | | | | |
| 23 | sub-sub-class | | **Tutorial** | | | | | | | |
| 24 | sub-sub-class | | **Workshop** | | | | | | | |
| 25 | super-class | | **Organization** | hasCity | **City** | | | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 26 | | | **Organization** | hasCountry | **Country** | | | | | |
| 27 | sub-class | | **Company** | | | | | | | |
| 28 | sub-class | | **University** | | | | | | | |
| 29 | super-class | | **Person** | employedBy | **Organization** | | | | | |
| 30 | | | **Person** | hasCity | **City** | | | | | |
| 31 | | | **Person** | hasCountry | **Country** | | | | | |
| 32 | sub-class | | **Administrator** | | | | | | | |
| 33 | sub-class | | **Assistant** | | | | | | | |
| 34 | sub-class | | **Author** | writes | **Contribution** | | | | | |
| 35 | sub-class | | **Chair_PC** | | | | | | | |
| 36 | sub-class | | **Member_PC** | expertOn | **Topic** | | | | | |
| 37 | | | **Member_PC** | reviews | **Contribution** | | | | | |
| 38 | sub-class | | **Participant** | | | | | | | |
| 39 | sub-sub-class | | **Member** | | | | | | | |
| 40 | sub-sub-class | | **Regular** | | | | | | | |
| 41 | sub-sub-class | | **Student** | | | | | | | |
| 42 | sub-class | | **Scholar** | studyAt | **University** | | | | | |
| 43 | sub-class | | **Science_Worker** | | | | | | | |
| 44 | sub-class | | **Volunteer** | | | | | | | |
| 45 | super-class | | **Topic** | | | | | | | |

Evaluation of the method align++

**Ontology B** **crs_dr** **domain of interest:** tool for conference organization support

**Modeling focus on:** *authors and their papers* **owl:Classes** 14

**owl:ObjectProperties** 15

**iweigthing Importance Degrees** related to the predefined modeling focus

| ID | IS-A Taxonomie | inverse OP | Domain | ObjectProperty | Range | Lowest Importance | Low Importance | Middle Importance | High Importance | Highest Importance |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | super-class | | **document** | | | | | | | |
| 3 | sub-class | inverse to ID 5 | **abstract** | part_of_article | **article** | | | | | |
| 4 | sub-class | inverse to ID 14 | **article** | article_written_by | **author** | | | | | |
| 5 | | inverse to ID 3 | **article** | has_abstract | **abstract** | | | | | |
| 6 | | | **article** | has_author | **author** | | | | | |
| 7 | | | **article** | has_reviewer | **reviewer** | | | | | |
| 8 | sub-class | inverse to ID 19 | **review** | review_written_by | **reviewer** | | | | | |
| 9 | super-class | | **event** | | | | | | | |
| 10 | sub-class | | **conference** | has_program | **program** | | | | | |
| 11 | sub-class | | **pc_meeting** | | | | | | | |
| 12 | sub-class | | **session** | | | | | | | |
| 13 | super-class | | **person** | | | | | | | |
| 14 | sub-class | inverse to ID 4 | **author** | writes_article | **article** | | | | | |
| 15 | | | **author** | assigns_article_to_conference | **article** | | | | | |
| 16 | sub-class | | **chair** | assigns_reviewers_to_article | **reviewer** | | | | | |
| 17 | | inverse to ID 20 | **chair** | creates_program | **program** | | | | | |
| 18 | sub-class | | **participant** | submits_to_conference | **conference** | | | | | |
| 19 | sub-class | inverse to ID 8 | **reviewer** | writes_review | **review** | | | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | super-class | inverse to ID 17 | **program** | program_created_by | **chair** | | | | | |
| 21 | | | **program** | has_session | **session** | | | | | |

Evaluation of the method align++

**Ontology A** **confOf** **domain of interest:** tool for conference organization support

**Modeling focus on:** *Events and Organizations* **owl:Classes** 38

**owl:ObjectProperties** 13

**iweigthing Importance Degrees** related to the predefined modeling focus

| ID | IS-A Taxonomie | inverse OP | Domain | ObjectProperty | Range | Lowest Importance | Low Importance | Middle Importance | High Importance | Highest Importance |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | super-class | | **City** | | | | | | | |
| 2 | super-class | | **Contribution** | dealsWith | **Topic** | | | | | |
| 3 | | inverse to 34 | **Contribution** | writtenBy | **Author** | | | | | |
| 4 | sub-class | | **Paper** | | | | | | | |
| 5 | sub-class | | **Poster** | | | | | | | |
| 6 | sub-class | | **Short_paper** | | | | | | | |
| 7 | super-class | | **Country** | | | | | | | |
| 8 | super-class | | **Event** | | | | | | | |
| 9 | sub-class | | **Administrative_event** | follows | **Administrative_event** | | | | | |
| 10 | | | **Administrative_event** | parallel_with | **Administrative_event** | | | | | |
| 11 | sub-sub-class | | **Camera_Ready_event** | | | | | | | |
| 12 | sub-sub-class | | **Registration_of_participants_event** | | | | | | | |
| 13 | sub-sub-class | | **Reviewing_event** | | | | | | | |
| 14 | sub-sub-class | | **Reviewing_results_event** | | | | | | | |
| 15 | sub-sub-class | | **Submission_event** | | | | | | | |
| 16 | sub-class | | **Social_event** | | | | | | | |
| 17 | sub-sub-class | | **Banquet** | | | | | | | |
| 18 | sub-sub-class | | **Reception** | | | | | | | |
| 19 | sub-sub-class | | **Trip** | | | | | | | |
| 20 | sub-class | | **Working_event** | hasAdminEvent | **Administrative_event** | | | | | |
| 21 | | | **Working_event** | hasTopic | **Topic** | | | | | |
| 22 | sub-sub-class | | **Conference** | | | | | | | |
| 23 | sub-sub-class | | **Tutorial** | | | | | | | |
| 24 | sub-sub-class | | **Workshop** | | | | | | | |
| 25 | super-class | | **Organization** | hasCity | **City** | | | | | |
| 26 | | | **Organization** | hasCountry | **Country** | | | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 27 | sub-class | | **Company** | | | | | | | | |
| 28 | sub-class | | **University** | | | | | | | | |
| 29 | super-class | | **Person** | employedBy | **Organization** | | | | | | |
| 30 | | | **Person** | hasCity | **City** | | | | | | |
| 31 | | | **Person** | hasCountry | **Country** | | | | | | |
| 32 | sub-class | | **Administrator** | | | | | | | | |
| 33 | sub-class | | **Assistant** | | | | | | | | |
| 34 | sub-class | | **Author** | writes | **Contribution** | | | | | | |
| 35 | sub-class | | **Chair_PC** | | | | | | | | |
| 36 | sub-class | | **Member_PC** | expertOn | **Topic** | | | | | | |
| 37 | | | **Member_PC** | reviews | **Contribution** | | | | | | |
| 38 | sub-class | | **Participant** | | | | | | | | |
| 39 | sub-sub-class | | **Member** | | | | | | | | |
| 40 | sub-sub-class | | **Regular** | | | | | | | | |
| 41 | sub-sub-class | | **Student** | | | | | | | | |
| 42 | sub-class | | **Scholar** | studyAt | **University** | | | | | | |
| 43 | sub-class | | **Science_Worker** | | | | | | | | |
| 44 | sub-class | | **Volunteer** | | | | | | | | |
| 45 | super-class | | **Topic** | | | | | | | | |

APPENDIX C

# Paired t-test in R

*Contribution* ($O_A$) and *article* ($O_B$) in `SCENARIO 1`
*Contribution* ($O_A$) and *article* ($O_B$) in `SCENARIO 2`

Transcript of a R session.

```
# SCENARIO 1

> Contribution_Sc1
 [1] 0.95 0.95 0.95 0.95 0.85 0.85 0.85 0.85 0.85 0.85 0.95 0.95 0.95 0.95 0.95
[16] 0.95 0.95 0.95
> article_Sc1
 [1] 0.90 0.90 0.85 0.90 0.95 0.95 0.95 0.80 0.85 0.85 0.95 0.85 0.90 0.95 0.95
[16] 0.95 0.95 0.90

> diff_Sc1<-(Contribution_Sc1-article_Sc1)
> summary(diff_Sc1)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.10000  0.00000  0.00000  0.01111  0.05000  0.10000

> t.test(Contribution_Sc1,article_Sc1,paired=T)

        Paired t-test

data:  Contribution_Sc1 and article_Sc1
t = 0.7757, df = 17, p-value = 0.4486    # p > alpha/2 is not significant
                                         # H0 cannot be rejected
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01910835  0.04133057
sample estimates:
mean of the differences
             0.01111111



## SCENARIO 2

> Contribution_Sc2
```

```
 [1] 0.05 0.15 0.15 0.15 0.05 0.15 0.05 0.15 0.15 0.05 0.15 0.15 0.05 0.05 0.15
[16] 0.05 0.05 0.15

> diff_Sc2<-(article_Sc1-Contribution_Sc2)
> summary(diff_Sc2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6500  0.7500  0.8000  0.8000  0.8875  0.9000

> t.test(article_Sc1,Contribution_Sc2,paired=T)

        Paired t-test

data:  article_Sc1 and Contribution_Sc2
t = 41.2669, df = 17, p-value < 2.2e-16    # p < alpha/2 is significant
                                           # H0 is rejected in favour of H1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7590991 0.8409009
sample estimates:
mean of the differences
                    0.8

```

# Bibliography


Neil M. Agnew, Kenneth M. Ford, and Patrick J. Hayes. Expertise In Context: Personally Constructed, Socially Selected and Reality-Relevant? *International Journal of Expert Systems*, 7 (1):65–88, 1993. available at `http://www.psych.yorku.ca/agnew/documents/AgnewFord_Hayes.pdf` (checked online January-20-2011).

Kent Bach. Semantic, Pragmatic. *In: J. Keim Campbell, M. O'Rourke, and D. Shier, eds., Meaning and Truth, New York: Seven Bridges Press*, pages 284–292, 2002. available at `http://userwww.sfsu.edu/~kbach/SPD.htm` (checked online December-20-2010).

Massimo Benerecetti, Paolo Bouquet, and Chiara Ghidini. Formalizing Belief Reports – The Approach and a Case Study. In *Proceedings AIMSA'98, 8th International Conference on Artificial Intelligence, Methodology, Systems, and Applications*, Sozopol (BG), September 1998. available at `http://dx.doi.org/10.1007/BFb0057435` (checked online July-20-2010).

Massimo Benerecetti, Paolo Bouquet, and Chiara Ghidini. Contextual Reasoning Distilled. *In: Journal of Theoretical and Experimental Artificial Intelligence, New York: Seven Bridges Press*, pages 279–305, 2000. available at `http://people.na.infn.it/~bene/pub.html` (checked online December-15-2010).

Massimo Benerecetti, Paolo Bouquet, and Chiara Ghidini. On the dimensions of context dependence. In *Third International and Interdisciplinary Conference, CONTEXT*, Dundee (UK), July 2001. available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.81.5050` (checked online July-20-2010).

Ana Belen Benitez, Di Zhong, Shih-Fu Chang, and John R. Smith. MPEG-7 MDS Content Description Tools and Applications. In *Proceedings of the 9th International Conference on Computer Analysis of Images and Patterns, CAIP'01*, volume 2124 of *Lecture Notes in Computer Science*, pages 41–52, London (UK), 2001. Springer Verlag. available at `http://www.springerlink.com/content/g4crecnldq1cprq4/` (checked online March-12-2010).

Richard V. Benjamins and Dieter Fensel. The ontological engineering initiative (ka)2. In Nicola Guarino, editor, *Proceedings of the First International Conference (FOIS'98)*, Formal Ontology in Information Systems, Trento (IT), June 1998. IOS Press.

Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possiblities. *Scientific American*, 284(5):34–43, May 2001.

Ted James Biggerstaff and C. Richter. Reusability Framework, Assessment, and Directions. *IEEE Software*, 4(2):10.1145/73103.73104, March 1987. available at `http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1695709` (checked online October-02-2010).

Elena Paslaru Bontas. Using Context Information to Improve Ontology Reuse. In *Doctoral Workshop at the 17th Conference on Advanced Information System Engineering, CAISE'05*, Berlin (DE), 2005. Freie Universität Berlin, Institut für Informatik. available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.6971` (checked online July-01-2010).

Elena Paslaru Bontas-Simperl and Christoph Tempich. Ontology Engineering: A Reality Check. In *OTM Conferences*, volume 4275 of *Lecture Notes in Computer Science*, pages 836–854, Montpellier (FR), 2006. Springer Verlag. available at `http://dx.doi.org/10.1007/11914853_51` (checked online September-02-2010).

Paolo Bouquet and Luciano Serafini. Context and Contextual Reasoning. 12th European Summer School in Logic, Language and Information, Birmingham (UK), August 2000. `http://sra.itc.it/people/serafini/esslli-course-on-contextual-reasoning.html` (checked online August-10-2010).

Paolo Bouquet, Antonia Doná, Luciano Serafini, and Stefano Zanobini. ConTeXtualized local ontology specification via CTXML. In *18th International Conference on Principles of Knowledge Representation and Reasoning, AAAI2002*, Edmonton, Alberta (CA), July 2002. available at `http://dit.unitn.it/~bouquet/papers/AAAI2002-MN.pdf` (checked online July-05-2010).

Paolo Bouquet, Fausto Giunchiglia, Frank von Harmelen, Luciano Serafini, and Heiner Stuckenschmidt. C-OWL: Contextualizing Ontologies. In *Journal Of Web Semantics*, pages 164–179. Springer Verlag, 2003a. available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.6680` (checked online March-03-2010).

Paolo Bouquet, Bernardo Magnini, Luciano Serafini, and Stefano Zanobini. A SAT-Based Algorithm For Context Matching. Technical Report 0306-10, ITC-IRST Centro per la Ricerca Scientifica e Technologica, 38050 Povo, Trento (IT), June 2003b. available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.6557` (checked online July-04-2010).

Paolo Bouquet, Jérôme Euzenat, Enrico Franconi, Luciano Serafini, Giorgos Stamou, and Sergio Tessaris. *D2.2.1: Specification of a common framework for characterizing alignment*. Knowledge Web Consortium, August 2004. available at `http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D2.2.1v2.pdf` (checked online June-30-2010).

Frank Budinsky, David Steinberg, Ed Merks, Raymond Ellersick, and Timothy J. Grose. *eclipse Modeling Framework*. Addison-Wesley Fachbuchverlag, 2004.

Rudolf Carnap. Logical Foundations of Probability, 1950. available at `http://evans-experientialism.freewebspace.com/carnap03.htm` (checked online February-15-2011).

Robyn Carston. The Semantics/Pragmatics Distinction: A View from Relevance Theory. In *UCL WORKING PAPERS IN LINGUISTICS*, volume 7, pages 1–30. Elsevier Science, 1998. available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.5184` (checked online July-28-2010).

Robyn Carston. *Pragmatics and Linguistic Underdeterminacy*, chapter 1, pages 1–79. Wiley Online Library, January 2008. available at `http://onlinelibrary.wiley.com/doi/10.1002/9780470754603.ch2/summary` (checked online December-08-2010).

Hans Chalupsky. OntoMorph: A translation system for symbolic logic. In *7th International Conference on Principles of Knowledge Representation and Reasoning, KR2000*, Breckenridge (CO US), April 2000. available at `www.isi.edu/~hans/publications/KR2000.ps` (checked online March-28-2010).

Noam Chomsky. *Rules and Representation*. Columbia University Press, New York (NY US), 1980.

Andy Clark. *Mindware. An Introduction to the Philosophy of Cognitive Science*. Oxford University Press, New York (NY US), 2001. available at `http://www.scribd.com/doc/8680720/Andy-Clark-Mindware-An-Introduction-to-the-Philosophy-of-Cognitive-Science-2001` (checked online October-18-2010).

Jos de Bruijn, Marc Ehrig, Cristina Feier, Francisco Martín-Recuerda, Francois Scharffe, and Moritz Weiten. *Ontology mediation, merging and aligning*. Semantic Web Technologies: Trends and Research in Ontology-based Systems. John Wiley & Sons, Ltd., July 2006. available at `http://www.dit.unitn.it/~p2p/RelatedWork/Matching/mediation-chapter.pdf` (checked online May-09-2010).

Klaus Dethloff. Prädikatenlogik der ersten Stufe mit Identität, July 2001.

Dubravko Dolić. *Statistik mit R*. Olde Verlag, 2004.

Rudolf Dutter. *Statistik und Wahrscheinlichkeitstheorie*. Institut für Statistik und Wahrscheinlichkeitstheorie, Vienna University of Technology, September 2002.

Marc Ehrig. *Ontology Alignment: Bridging the Semantic Gap*, volume 4 of *Semantic Web And Beyond Computing for Human Experience*. Springer Verlag, 1st edition, 2007. ISBN 978-0-387-36501-5.

Marc Ehrig and Steffen Staab. QOM – Quick Ontology Mapping. In *Proceedings of the Third International Semantic Web Conference*, volume 3298. Springer Verlag, 2004. available at http://www.aifb.kit.edu/web/Inproceedings701 (checked online July-15-2010).

Marc Ehrig and York Sure. Ontology Mapping - An Integrated Approach. In *1th European Semantic Web Symposium, ESWS 2004*, volume 3053 of *Lecture Nodes in Computer Science*, pages 76–91, Heraklion (GR), 2004. Springer Verlag. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.4772 (checked online August-11-2010).

Marc Ehrig and York Sure. FOAM: Framework for Ontology Alignment and Mapping; results of the ontology alignment initiative. In *Proceedings of the Workshop on Integrating Ontologies*, pages 72–76. CEUR-WS.org, 2005. available at http://www.aifb.kit.edu/web/Inproceedings1111 (checked online August-10-2010).

Marc Ehrig, Peter Haase, Mark Hefke, and Nenad Stojanovic. Similarity for Ontologies - a Comprehensive Framework. In *In Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, at PAKM 2004*, Vienna (AT), 2004. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.4494 (checked online July-01-2010).

Roand Eller, Walter S. A. Schwaiger, and Richard Federa. *Bankenbezogene Risiko- und Erfolgsrechnung*. Schäffer-Poeschel Verlag, Stuttgart (DE), 2002.

Jeffrey L. Elman. Representation and Structure in Connectionist Models. Technical Report 8903, Center for Research in Language, University of California, San Diego (CA US), August 1989. available at http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA259504&Location=U2&doc=GetTRDoc.pdf (checked online October-04-2010).

Herbert B. Enderton. *A Mathematical Introduction to Logic, Second Edition*. Harcourt/Academic Press, San Diego (CA US), second edition, 1972.

Neil A. Ernst, Margaret-Anne Storey, and Polly Allen. Cognitive Support for Ontology Modeling. *International Journal of Human-Computer Studies*, 62(5):10.1016.j.ijhcs.2005.02.006, May 2005. available at http://fink08.files.wordpress.com/2009/12/ijhcs-protege.pdf (checked online September-13-2010).

Jérôme Euzenat. Towards Formal Knowledge Intelligibility at the Semiotic Level. In *ECAI Workshop on Applied Semiotics: Control Problems*, Berlin (DE), 2000. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.2753 (checked online June-15-2010).

Jérôme Euzenat. Towards a Principled Approach to Semantic Interoperability. In *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle (WA US), 2001. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.9779 (checked online June-27-2010).

Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer Verlag, Heidelberg (DE), 2007.

Jérôme Euzenat and Petko Valtchev. An integrative proximity measure for ontology alignment, 2003. available at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-82/SI_paper_06.pdf (checked online October-27-2009).

Jérôme Euzenat and Petko Valtchev. Similarity-based ontology alignment in OWL-Lite. In *The 16th European Conference on Artificial Intelligence, ECAI-04*, Valencia (ES), 2004. available at http://www.iro.umontreal.ca/~owlola/pdf/align-ECAI04-FSub.pdf (checked online March-05-2010).

Jérôme Euzenat, Thanh Le Bach, Jesús Barrasa, Paolo Bouquet, Jan De Bo, Rose Dieng, Marc Ehrig, Manfred Hauswirth, Mustafa Jarrar, Ruben Lara, Diana Maynard, Amedeo Napoli, Giorgos Stamou, Heiner Stuckenschmidt, Pavel Shvaiko, Sergio Tessaris, Sven Van Acker, and Ilya Zaihrayeu. *D2.2.3: State of the art on ontology alignment*. KnowledgeWeb Consortium, August 2004. available at http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D2.2.3.pdf (checked online June-30-2010).

Jérôme Euzenat, Raúl García Castro, and Marc Ehrig. *D2.2.2: Specification of a benchmarking methodology for alignment techniques*. Knowledge Web Consortium, February 2005. available at http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D2.2.2.pdf (checked online June-30-2010).

Jérôme Euzenat, Francois Scharffe, and Luciano Serafini. *D2.2.6: Specification of the delivery alignment format*. KnowledgeWeb Consortium, February 2006. available at http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D2.2.6.pdf (checked online July-08-2010).

Ludwig Fahrmeir, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik*. Springer Verlag, 4 edition, 2003.

Sean M. Falconer. Cognitive Support for Human-Guided Mapping Systems. Technical report, University of Victoria, 2007. available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.8724 (checked online September-09-2010).

Sean M. Falconer and Margaret-Anne Storey. A cognitive support framework for ontology mapping. In *6th International The semantic web and 2nd Asian conference on Asian semantic web conference*, Busan (KR), 2007. available at http://portal.acm.org/citation.cfm?id=1785172 (checked online July-17-2010).

Sean M. Falconer, Natalya Fridman Noy, and Margaret-Anne Storey. Towards understanding the needs of cognitive support for ontology mapping. In *International Workshop on Ontology Matching, OM-2006*, Georgia (GA US), November 2006. available at http://www.dit.unitn.it/~p2p/OM-2006/3-Falconer-TP-OM%2706.pdf (checked online April-21-2010).

Sean M. Falconer, Natalya F. Noy, and Margaret-Anne Storey. Ontology Mapping - A User Survey. In *The 2nd International Workshop Ontology Matching, OM 07*, Busan (KR), 2007. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.5740 (checked online April-17-2010).

Anita Fetzer. *Recontextualizing Context: Grammaticality meets Appropriateness*. John Benjamins B. V., 2004. available at http://ebookee.org/Anita-Fetzer-Recontextualizing-Context-Grammaticality-Meets-Appropriateness_406513.html (checked online October-17-2010).

Peter Filzmoser. *Multivariate Statistik*. Institut für Statistik und Wahrscheinlichkeitstheorie, Vienna University of Technology, January 2003.

Jürgen Franke, Wolfgang Härdle, and Christian Hafner. *Einführung in die Statistik der Finanzmärkte*, volume 2 of *Statistik und ihre Anwendungen*. Springer Verlag, 1st edition, 2004.

Fabien Gandon. Ontology Engineering: a Survey and a Return on Experience. Technical Report 4396, Institut de Recherche en Informatique et Automatique (INRIA), March 2002. available at http://hal.inria.fr/docs/00/07/21/92/PDF/RR-4396.pdf (online checked June-18-2010).

Dragan Gašević, Dragan Djurić, and Vladan Devedžić. *Model Driven Architecture and Ontology Development*. Springer Verlag, 2006.

Chiara Ghidini and Fausto Giunchiglia. Local Models Semantics, or Contextual Reasoning = Locality + Compatibility. *Aritificial Intelligence, AIJ01*, 127(2):10.1016/S0004-3702(01)00064-9, January 2001. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.7598 (checked online July-04-2010).

Fausto Giunchiglia. Contextual Reasoning. *Epistemologia, special issue on I Linguaggi e le Macchine*, 345:345–364, 1992. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.7615 (checked online July-05-2010).

Fausto Giunchiglia and Paolo Bouquet. Introduction to contextual reasoning. An Artificial Intelligence perspective. In B. Kokinov, editor, *Perspectives on Cognitive Science*, volume 3, pages 138–159, Sofia (BG), 1997. NBU Press. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.83.6735 (online checked June-08-2010).

Fausto Giunchiglia and Luciano Serafini. Multilanguage Hierarchical Logics, or: How we can do without modal logics. *Aritificial Intelligence*, 65(1):29–70, 1994. available at http://disi.unitn.it/~context/paper/GS91a.pdf (online checked August-08-2010).

Fausto Giunchiglia and Pavel Shvaiko. SEMANTIC MATCHING. Technical Report DIT-03-013, University of Trento, Department of Information and Communication Thechnology, 38050 Povo, Trento (IT), Via Sommarive 14, April 2003. available at http://portal.acm.org/citation.cfm?id=991811 (online checked June-08-2010).

Fausto Giunchiglia and Mikalai Yatskevich. ELEMENT LEVEL SEMANTIC MATCHING. Technical Report DIT-04-035, University of Trento, Department of Information and Communication Thechnology, 38050 Povo, Trento (IT), Via Sommarive 14, June 2004. available at http://eprints.biblio.unitn.it/archive/00000591/ (online checked March-21-2010).

Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. Discovering Missing Background Knowledge in Ontology Matching. Technical Report DIT-06-005, University of Trento Department of Information and Communication Thechnology, 38050 Povo, Trento (IT), Via Sommarive 14, February 2006. available at http://eprints.biblio.unitn.it/archive/00000953/ (checked online June-07-2010).

Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering*. Springer Verlag, 2003.

Herbert Paul Grice. *Logic and Conversation*, volume 3 of *Speech acts, Syntax and Semantics*. Academic Press, NY US, 1975. available at http://www.mystfx.ca/academic/philosophy/Cook/2008-09/Grice-Logic.pdf (checked online October-10-2010).

Thomas R. Gruber. A Translation Approach to Portable Ontology Specifications. Technical Report KSL 92-71, Knowledge System Laboratory, Stanford University, 701 Welch Road, Building C Palo Alto (CA US), 94304, April 1993. available at http://tomgruber.org/writing/index.htm (checked online July-01-2010).

Thomas R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43(5–6):10.1006/ijhc.1995.1081, November/Dezember 1995. available at http://portal.acm.org/citation.cfm?id=219701 (checked online August-10-2010).

Michael Grüninger and Mark S. Fox. Methodology for the Design and Evaluation of Ontologies. In *International Joint Conference on Artificial Inteligence IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*, Toronto (CA), April 1995. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.8723 (checked online April-28-2010).

Nicola Guarino. Formal Ontology and Information Systems. In *1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, pages 3–15, Trento (IT), June 1998. IOS Press. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.1776 (checked online September-17-2010).

Ramanathan V. Guha, Rob McCool, and Richard Fikes. Contexts for the Semantic Web. In *International Semantic Web Conference, ISWC 2004*, volume 3298 of *Lecture Notes in Computer Science*, pages 32–46. Springer Verlag, 2004. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.2368 (checked online August-05-2010).

Adil Hameed, Alun Preece, and Derek Sleeman. Ontology Reconciliation. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 231–250. University of Aberdeen, Department of Computer Science, Springer Verlag, 2004. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.95.9827 (checked online March-01-2010).

Bernhard Haslhofer and Wolfgang Klas. A Survey of Techniques for Achieving Metadata Interoperability. *ACM Computing Surveys (CSUR)*, 42(2):10.1145/1667062.1667064, February 2010. available at http://portal.acm.org/citation.cfm?id=1667064&dl=ACM&coll=DL&CFID=14214136&CFTOKEN=56663786 (checked online March-10-2010).

Pascal Hitzler, Jérôme Euzenat, Markus Krötzsch, Luciano Serafini, Heiner Stuckenschmidt, HolgerWache, and Antoine Zimmermann. *D2.2.5: Integrated view and comparison of alignment semantics*. KnowledgeWeb Consortium, January 2006. available at http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D2.2.5.pdf (checked online July-08-2010).

Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph, and York Sure. *Semantic Web*. Springer Verlag, 1 edition, 2008.

Douglas Hofstadter. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Perseus Books Group, New York (NY US), 1995. available at http://www.questia.com/read/99872249 (checked online October-15-2010).

Matthew Horridge, Simon Jupp, Georgina Moulton, Alan Rector, Robert Stevens, and Chris Wroe. *A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools*. University of Manchester, 1.1 edition, October 2007. available at http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_1.pdf (checked online July-05-2010).

Todd C. Hughes and Benjamin C. Ashpole. The Semantics of Ontology Alignment. In *Proceedings of the 2004 Performance Metrics for Intelligent Systems Workshop (PerMIS'04),*, Gaithersburg (MD US), August 2004. available at http://www.atl.lmco.com/papers/1243.pdf (checked online November-11-2009).

Christian Janiesch. Situation vs. Context: Considerations on the Level of Detail in Modelling Method Adaptation. In *43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE Computer Society, 2010. ISBN 978-0-7695-3869-3. URL http://dblp.uni-trier.de/db/conf/hicss/hicss2010.html#Janiesch10. available at http://www.computer.org/portal/web/csdl/doi/10.1109/HICSS.2010.340 (checked online May-28-2010).

Mario Jeckle, Chris Rupp, Jürgen Hahn, Barbara Zengler, and Stefan Queins. *UML 2 glasklar*. Verlag Hanser, 3 edition, 2003.

Yannis Kalfoglou. EXPLORING ONTOLOGIES. In *Handbook of Software Engineering and Knowledge Engineering*, volume 1, pages 863–887, 80 South Bridge Edinburgh (SCO UK), EH1 1HN, 2000. School of Artificial Intelligence, University of Edinburgh. available at http://eprints.ecs.soton.ac.uk/10528/ (checked online May-17-2010).

Michel Klein. Combining and Relating Ontologies: An Analysis of Problems and Solutions. In A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt, and M. Uschold, editors, *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle (WA US), August 2001. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.9623 (checked online May-28-2010).

*The OWL Import Benchmark Suite. Formal description of the benchmarks*. Knowledge Web Consortium, 2007. URL http://knowledgeweb.semanticweb.org/benchmarking_interoperability/owl/. available at http://knowledgeweb.semanticweb.org/benchmarking_interoperability/owl/files/description.pdf (checked online May-15-2010).

Metaphysics Research Lab. Stanford Encyclopedia of Philosophy, 2006. URL http://plato.stanford.edu/entries/pragmatics/. checked online August-09-2010.

Monika Lanzenberger and Jennifer Sampson. AlViz - A Tool for Visual Ontology Alignment. In *10th International Conference on Information Visualisation*, pages 430–440, London (UK), 2006. IEEE Computer Society. available at http://www.ifs.tuwien.ac.at/node/12882 (checked online October-19-2009).

Monika Lanzenberger, Jennifer Sampson, Markus Rester, Yannick Naudet, and Thibaud Latour. Visual ontology alignment for knowledge sharing and reuse. *Journal of Knowledge Management*, 12(6):10.1108/13673270810913658, 2008. available at http://lpis.csd.auth.gr/mtpx/km/material/JKM-12-6h.pdf (checked online November-09-2009).

Odd Ivar Lindland, Guttorm Sindre, and Arne Sølvberg. Understanding Quality in Conceptual Modeling. *IEEE Software*, 11(2):10.1109/52.268955, March 1994. available at http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=268955 (checked online August-30-2010).

Alexander Maedche and Steffen Staab. Measuring Similarity between Ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, EKAW 2002*, volume 2473 of *Lecture Notes In Computer Science*, pages 251–263, Siguenza (ES), October 2002. Springer Verlag. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.4280 (checked online August-26-2010).

Bernardo Magnini, Luciano Serafini, and Manuela Speranza. Making Explicit the Semantics Hidden in Schema Models. In *ISWC 2003, Workshop on Human Language Technology for the Semantic Web and Web Services*, Sanibel Island (FL US), October 2003. available at http://tcc.itc.it/people/speranza/publications.html (checked online April-08-2010).

Jiří Matoušek and Jaroslav Nešetřil. *Diskrete Mathematik*. Springer Verlag, 2 edition, 2007.

Alexandra Mazak, Monika Lanzenberger, and Bernhard Schandl. Enhancing Structure-based Ontology Alignment by Enriching Models with Importance Weightings. In *3rd International Workshop on Ontology Alignment and Visualization, OnAV'10*, Krakow (PL), February 2010a. available at `http://publik.tuwien.ac.at/files/PubDat_187532.pdf` (checked online July-01-2010).

Alexandra Mazak, Bernhard Schandl, and Monika Lanzenberger. align++: A Heuristic-based Method for Approximating the Mismatch-at-Risk in Schema-based Ontology Alignment. In *International Conference on Knowledge Engineering and Ontology Development, KEOD 2010*, Valencia (ES), October 2010b. available at `http://www.informatik.univie.ac.at/publication.php?pid=7170` (checked online August-29-2010).

John McCarthy. Notes on Formalizing Context. In *13th International Joint Conference on Artificial Intelligence, IJCAI'93*, pages 555–560, Chambery (FR), 1993. Knowledge Representation. available at `http://www-formal.stanford.edu/jmc/context3/context3.html` (checked online June-29-2010).

David Meintrup and Stefan Schäffler. *Stochastik*. Statistik und ihre Anwendungen. Springer Verlag, 1st edition, 2005.

Donald A. Norman. *Things That Make Us Smart: Defending Human Attributes In The Age Of The Machine*. Perseus Books Group, 1993.

Donald A. Norman and Steven Draper. *User Centered System Design: New Perspectives on Human-Computer Interaction*. Lawrence Erlbaum Associates, Inc, 1986.

Natalya F. Noy. Ontology Mapping. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, volume 2 of *International Handbooks on Information Systems*, pages 573–590. Springer Verlag, 2009. available at `http://www.springerlink.com/content/j395h1m372776r80/` (checked online April-02-2010).

Natalya Fridman Noy. Semantic Integration: A Survey Of Ontology-Based Approaches. In *SIGMOD Record*, volume 33, New York (NY US), December 2004. ACM. available at `http://www.mendeley.com/research/semantic-integration-a-survey-of-ontologybased-approaches/` (checked online July-04-2010).

Natalya Fridman Noy and Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. Technical Report SMI-2001-0880, Stanford University, Stanford (CA US), 94305, March 2001. available at `http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html` (checked online March-30-2010).

Natalya Fridman Noy and Mark A. Musen. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pages

450–455, Austin (TX US), August 2000. AAAI Press. available at http://portal.acm.org/citation.cfm?id=647288.721118&coll=GUIDE&dl=GUIDE (checked online August-26-2010).

Natalya Fridman Noy and Mark A. Musen. Anchor-PROMPT: Using Non-local Context for Semantic Matching. In *Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, Seattle (WA US), 2001. available at http://bmir.stanford.edu/publications/view.php/anchor_prompt_using_non_local_context_for_semantic_matching (checked online March-02-2010).

Natalya Fridman Noy and Mark A. Musen. Evaluation Ontology Mapping Tools: Requirements and Experience. In *Workshop on Evaluation of Ontology Tools at the 13th International Conference on Knowledge Engineering and Knowledge Management, EON2002*, Siguenza (ES), 2002. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.3.8572 (checked online July-30-2010).

Natalya Fridman Noy and Mark A. Musen. The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping. *International Journal of Human-Computer Studies*, 59(6):10.1016/j.ijhcs.2003.08.002, December 2003. available at http://bmir.stanford.edu/publications/view.php/the_prompt_suite_interactive_tools_for_ontology_merging_and_mapping (checked online July-25-2010).

*Meta Object Facility (MOF) Core Specification*. OMG Object Management Group, 2.0 edition, January 2006. available at http://www.omg.org/spec/MOF/2.0/ (checked online July-11-2011).

*Ontology Definition Metamodel (ODM)*. OMG Object Management Group, 1.0 edition, May 2009. available at http://www.omg.org/spec/ODM/1.0/ (checked online July-11-2011).

Thomas Ottmann and Peter Widmayer. *Algorithmen und Datenstrukturen*. Spektrum Akademischer Verlag, 4 edition, 2002.

Aris M. Ouksel and Amit P. Sheth. Semantic Interoperability in Global Information Systems: A Brief Introduction to the Research Area and the Special Section. *SIGMOD Record*, 28(1): 10.1145/309844.309849, March 1999. available at http://doi.acm.org/10.1145/309844.309849 (checked online September-28-2010).

Eunhee Park and Han G. Woo. Reuse of Process Knowledge in Enterprise Systems Development. In *40th Annual Hawaii International Conference on System Sciences, HICSS'07*, Hawaii (HI US), January 2007. IEEE Computer Society. available at http://www.computer.org/portal/web/csdl/doi/10.1109/HICSS.2007.466 (checked online November-11-2010).

Thomas B. Passin. *Explorer's Guide to the Semantic Web*. Manning Publications Co., 2004.

Geert Poels, Ann Maes, Frederik Gailly, and Roland Paemeleire. Measuring the Perceived Semantic Quality of Information Models. In *ER 2005 Workshops AOIS, BP-UML, CoMoGIS, eCOMO, and QoIS*, Klagenfurt (AT), October 2005. available at http://www.springerlink.com/content/743t677jh4726546/ (checked online February-20-2011).

Christopher Potts. Formal Pragmatics. In Louise Cummings, editor, *The Routledge Encyclopedia of Pragmatics*, pages 167–170. Routledge, London, 2010. available at http://www.stanford.edu/~cgpotts/entries/potts-routledge08-formal-pragmatics.pdf (checked online November-26-2010).

Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal The International Journal on Very Large Data Bases*, 10(4):10.1007/s007780100057, Dezember 2001. available at http://dbs.uni-leipzig.de/de/publication/title/a_survey_of_approaches_to_automatic_schema_matching (checked online November-04-2009).

Erhard Rahm, Hong-Hai Do, and Sabine Maßmann. Matching Large XML Schemas. In *SIGMOD Record*, volume 33, New York (NY US), December 2004. ACM. available at http://www.mendeley.com/research-papers/search/?query=Matching+Large+XML+Schemas&x=0&y=0 (checked online March-26-2010).

Balasubramaniam Ramesh and Vasant Dhar. Supporting Systems Development by Capturing Deliberations During Requirements Engineering. *IEEE Transactions on Software Engineering*, 18(6):10.1109/32.142872, June 1992. available at http://portal.acm.org/citation.cfm?id=129967 (checked online October-17-2010).

Jennifer Sampson. *A comprehensive framework for ontology alignment quality*. PhD thesis, Norwegian University of Science and Technology, Trondheim (NO), March 2007. available at http://www.idi.ntnu.no/research/doctor_theses/sampsonj.pdf (checked online September-03-2010).

John Godfrey Saxe. The Blind Men and the Elephant, 1887. URL http://www.wordfocus.com/word-act-blindmen.html. available at http://www.wordfocus.com/word-act-blindmen.html (checked online June-17-2010).

Pavel Shvaiko and Jérôme Euzenat. A Survey of Schema-based Matching Approaches. Technical Report DIT-04-087, University of Trento, Department of Information and Communication Technology, October 2004. available at http://www.dit.unitn.it/~p2p/RelatedWork/Matching/JoDS-IV-2005_SurveyMatching-SE.pdf (checked online June-25-2010).

Elena Simperl. Reusing Ontologies on the Semantic Web: A feasibility study. *Elsevier*, 68(10): 905–925, 2009. available at http://dl.acm.org/citation.cfm?id=1598334 (checked online September-03-2010).

Paul R. Smart and Paula C. Engelbrecht. An Analysis of the Origin of Ontology Mismatches on the Semantic Web. In Aldo Gangemi and Jérôme Euzenat, editors, *Knowledge Engineering: Practice and Patterns 16th International Conference, EKAW 2008*, pages 120–135, Acitrezza (IT), 2008. Springer Verlag. available at `http://eprints.ecs.soton.ac.uk/15748/1/ITA_Research_Paper_EKAW2008_1v7.pdf` (checked online July-02-2010).

Dan Sperber and Deirdre Wilson. *Relevance: communication and cognition*. Blackwell Verlag, second edition, 1995.

Steffen Staab and Rudi Studer. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer Verlag, 2004.

Steffen Staab, Rudi Studer, Hans-Peter Schnurr, and York Sure. Knowledge Processes and Ontologies. *IEEE Intelligent Systems*, 16(1):10.1109/5254.912382, January 2001. available at `http://www.aifb.kit.edu/web/Article475` (checked online September-07-2010).

Werner A. Stahel. *Statistische Datenanalyse*. Vieweg & Sohn Verlagsgesellschaft mbH, 3rd edition, 2000.

Nenad Stojanovic. *Semantic Query Expansion*. PhD thesis, Universität Karlsruhe (TH), Institut AIFB, 2005.

Andreas Tolk. What Comes After the Semantic Web - PADS Implications for the Dynamic Web. In *20th Workshop on Principles of Advanced and Distributed Simulation, PADS'06*, Singapore (SG), May 2006. IEEE Computer Society. available at `http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1630709` (checked online September-27-2010).

Charles Travis. Pragmatics. *A Companion to the Philosophy of Language, Basil Blackwell Oxford*, pages 87–107, 1997.

Mike Uschold and Michael Grüninger. Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, 11(2):96–137, June 1996. available at `http://stl.mie.utoronto.ca/publications/ker.pdf` (checked online March-30-2010).

Mike Uschold, Mike Healy, Keith Williamson, Peter Clark, and Steven Woods. Ontology Reuse and Application. In *1st International Conference on Formal Ontology and Information Systems, FOIS'98*, pages 179–192, Trento (IT), June 1998. IOS Press. available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.3839` (checked online September-08-2010).

Teun A. van Dijk. Relevance in Text and Context: Discussion of Joseph E. Grimes' Preprint "Context Structure Patterns". In S. Allén, editor, *Proceedings of the Nobel Symposium*, volume 51 of *Proceedings of the Nobel Symposium*. available at `http://www.discourses.org/download/articles/` (checked online October-19-2010), Almquist and Wiksell, 1982.

Pepijn R. S. Visser, Dean M. Jones, T.J.M Bench-Capon, and M.J.R. Shave. An analysis of Ontology Mismatches; Heterogeneity versus Interoperability. In *AAAI 1997, Spring Symposium on Ontological Engineering*, Stanford (CA US), 1997. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.6709 (checked online June-22-2010).

Pepijn R.S. Visser and Zhan Cui. On Accepting Heterogeneous Ontologies in Distributed Architectures. In *13th European Conference on Artificial Intelligence, ECAI98*, Brighton (UK), August 1998. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.3541 (checked online July-02-2010).

Denny Vrandečić, Johanna Völker, Peter Haase, Thanh Tran Duc, and Philipp Cimiano. A Metamodel for Annotations of Ontology Elements in OWL DL. In *Proceedings of the 2nd Workshop on Ontologies and Meta-Modeling*, Karlsruhe (DE), October 2006. GI Gesellschaft für Informatik, York Sure and Saartje Brockmans and Jürgen Jung. available at (checked online May-09-2010).

*Resource Description Framework*. W3C, World Wide Web Consortium, February 1999. available at http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/ (checked online June-03-2010).

*OWL Web Ontology Language Guide*. W3C, World Wide Web Consortium, February 2004. available at http://www.w3.org/TR/owl-guide/ (checked online May-29-2010).

*OWL 2 Web Ontology Language*. W3C, World Wide Web Consortium, October 2009. available at http://www.w3.org/TR/owl2-overview/ (checked online March-09-2011).

Dennis Wagelaar. Contextual constraints in configuration languages. available at http://soft.vub.ac.be/svpp08/files/svpp08-wagelaar.pdf (checked online September-11-2010), 2008.

Andrew Walenstein. *Cognitive Support in Software Engineering Tools: A Distributed Cognition Framework*. PhD thesis, Simon Fraser University, 2002. available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.1911 (checked online September-09-2010).

Yair Wand and Ron Weber. Research Commentary: Information Systems and Conceptual Modeling–A Research Agenda. *Information Systems Research*, 13(4): 10.1287/isre.13.4.363.69, December 2002. available at http://portal.acm.org/citation.cfm?id=769474 (checked online September-04-2010).

Gang Wu, Juanzi Li, Ling Feng, and Kehong Wang. Identifying Potentially Important Concepts and Relations in an Ontology. In *Proceedings of the 7th International Conference on The Semantic Web*, Karlsruhe (DE), 2008. available at http://portal.acm.org/citation.cfm?id=1483155.1483159 (checked online May-29-2010).

Stefano Zanobini. *Semantic coordination: the model and an application to schema matching*. PhD thesis, International Doctorate School in Information and Communication Technology, University of Trento, Trento (IT), March 2006.

Alexandra Mazak
Dipl.-Ing. Mag.
Lerchenfelderstraße 124-126/1/14 0664/75052744
A-1080 Wien mazak@ifs.tuwien.ac.at

# Curriculum vitae

## Persönliche Angaben

| | |
|---|---|
| Name: | Alexandra Mazak, Dipl.-Ing. Mag. |
| Wohnhaft in: | Lerchenfelderstraße 124-126/1/14, 1080 Wien |
| Geburtsdatum: | 25.12.1969 |
| Geburtsort: | Wien |
| Familienstand: | ledig |

## Forschungsschwerpunkte

- Wissensbasierte Systeme: Entwicklung, Re-Use und Evaluierung von Ontologien
- Kognitive Engineering und Metakognition
- Systems Engineering: Anforderungsmanagement (Traceability)
- Linguistik: Pragmatik (Relevanztheorie)
- Stochastik

## Persönliche Fähigkeiten

Systemanalytisches, abstraktes und lösungsorientiertes Denkvermögen
Durchsetzungsvermögen
Selbständiges wissenschaftliches Arbeiten
Konsequente Arbeitsweise
Interdisziplinarität
Organisation und Strukturierung
Flexibilität und Eigeninitiative
Kommunikative Kompetenz und Teamfähigkeit

## Technische Fähigkeiten

Web Technologien (HTML, CSS, XML)
Semantic Web Technologien (RDF, OWL, SPARQL, Reasoning)
Datenbanksprachen (SQL)
Statistik Programme (R, S-Plus)
Diverse Frameworks (ANSI/ISA-95, ODM, MOF, EMF)

## Sprachkenntnisse

Englisch: in Wort und Schrift
Französisch: Grundkenntnisse

## Interessen

Sport (Bergsteigen, Skitourengehen), mein Hund Henry, Reisen

Alexandra Mazak
Dipl.-Ing. Mag.
Lerchenfelderstraße 124-126/1/14 0664/75052744
A-1080 Wien mazak@ifs.tuwien.ac.at

## Beruflicher Werdegang

| | |
|---|---|
| RSA Research Studios Austria Forschungsgesellschaft mbH, Wien | Jänner 2012<br>Senior Researcher im Bereich Cognitive Engineering<br>Forschungsschwerpunkt: die Einbindung kognitiver und meta-kognitiver Human Factors in der Softwaremodellierung. |
| SBA Research gGmbH<br>in Kooperation mit der Technischen Universität Wien | September 2010 bis Dezember 2011<br>Förderung der Dissertationsschrift im Zuge des Projekts FAMOS – Female Academy for Mentoring, Opportunities and Self-Development (Nachfolgeprojekt von FINCA). |
| Technische Universität Wien, Institut für Softwaretechnik und Interaktive Systeme | Februar 2009 bis August 2010<br>Assistentin im Projekt FINCA - Female Intrapreneurship Career Academy einem vom bmvit (Bundesministerium für Verkehr, Innovation und Technologie) durch die Programmlinie FEMtech Karrierewege finanzierten Kooperationsprojekts zwischen der Johannes Kepler Universität Linz, der Technischen Universität Wien und der Technischen Universität Graz, sowie 7 Partnerunternehmen.<br><br>Konzeption und Umsetzung eines Web-Portals zur interaktiven Vernetzung und Kooperation der wissenschaftlichen Partner, der Unternehmenspartner, sowie der Teilnehmerinnen im Projekt.<br>Einsatz von Web 2.0 Technologien zur dislozierten Vernetzung und zur Dokumentation von Verhaltensveränderungen der Teilnehmerinnen sowie anderer Projektergebnisse.<br>Betreuung der Projektteilnehmerinnen und Projektfirmen in Wien, sowie Erstellung des Zwischen- und Endberichts an die FFG. |
| P.Solutions Informations-technologien GmbH, Wien | März 2008 bis Juli 2008<br>Consulting<br>Leitung der Abteilung Innovation, Bereich Semantic IT am Ende der Laufzeit des Projekts SemDAV, eines von der FFG geförderten EU-Projektes zwischen P.Solutions, der Universität Wien und den Research Studios Austria (ARC Seibersdorf Research GmbH).<br><br>Entwicklung von semantischen Produkten basierend auf der SemDAV-Technologie. Konzeption und Modellierung einer Domain-Ontologie am konkret umgesetzten Beispiel eines semantischen Dokumentenmanagementsystems mit Erweiterungsmöglichkeiten zu einem semantischen Enterprise-Information-Management Systems. |
| Technische Universität Wien, Institut für Managementwissenschaften, Fachbereich Finanzwissenschaften und Controlling | Jänner 2007 bis Februar 2008<br>Universitätsassistentin<br><br>Lehrauftrag: VU Rechnungswesen 1 und 2<br>Forschungsschwerpunkte: Kybernetische Regelkreissysteme,<br>Föderal konzipiertes Controlling unter Unsicherheit<br>Entwicklung eines ERP-gestützten Controlling Systems<br>unter Einbindung der Time Driven Activity Based Costing (TD ABC) |

Alexandra Mazak
Dipl.-Ing. Mag.
Lerchenfelderstraße 124-126/1/14 0664/75052744
A-1080 Wien mazak@ifs.tuwien.ac.at

| | |
|---|---|
| | und des Standards ANSI/ISA-95 zur Abbildung von dezentralisierten Informationsflüssen. Implementierung eines Prototyps mittels einer objekt-relationalen Datenbank. |
| media-daten.at Onlineverlag GmbH | Firmengründung August 2004 bis August 2010<br>Eigentümerin 100% Anteile, Unternehmensführung<br>(Tochtergesellschaft der eventus Marketingservice GmbH, Umfirmierung zur wu-wei Onlineverlag GmbH)<br><br>Technische Konzeption und Implementierung eines relationalen Datenbanksystems zur Erfassung und Verwaltung von Informationen sämtlicher Werbemedien sowohl in technischem als auch preislichem Umfang. Die konzipierte Datenbank wurde technisch umgesetzt, implementiert und mittels eines Online Userinterface zur Nutzung fertiggestellt.<br>Konzeption und Umsetzung von Onlineplattformen für Tourismus und Gastronomie (www.austria-aktiv.at, www.garum.at). |
| eventus Marketingservice GmbH, Wien | Firmengründung Dezember 1994 bis August 2010<br>Eigentümerin Anteile 100%, Unternehmensführung<br><br>Tätigkeit in den Bereichen Direct Marketing, Database Marketing und Skitouren-Filme.<br>Entwicklung von Austria-aktiv.at einer interaktiven Plattform für Ski-, Rad-, Berg-, Wander- und Trekkingtouren. Alle markanten Punkte entlang einer Tour wurden als sogenannte Geonam-Objekte mit den entsprechenden GPS-Koordinaten referenziert. Es entstanden 150.000 Geoobjekte, die eine effiziente Anzeige und Visualisierung einer Tour gewährleisten. Entlang einer Tour kann zu über 1000 Hütten, Orten und dortigen Gastronomie- und Hoteleriebetrieben uvm. verzweigt werden. |
| go direct Marketingservice & HandelsgesmbH, Wien | Juli 1992 bis März 1995<br>Februar 1994, Übernahme 100%-Anteile an der go direct Marketingservice & HandelsgesmbH<br>Consulting, Unternehmensführung<br><br>Konzeption und Implementierung einer Marketingdatenbank für internationale Kunden (McDonalds, General Motors, OMV, IKEA) zur Planung, Steuerung und zur technischen Abwicklung von Direct Marketingmaßnahmen bzw. Kampagnen. |
| Werbeagentur RSCG Jasch & Schramm, Wien | Juli 1991 bis Juni 1992<br>Rezeptionistin, Assistentin |
| Kapsch AG, Wien | März 1990 bis Juni 1991<br>Sekretärin im Sondervertrieb,<br>Bereich : Private Kommunikationssysteme |

Alexandra Mazak
Dipl.-Ing. Mag.
Lerchenfelderstraße 124-126/1/14 0664/75052744
A-1080 Wien mazak@ifs.tuwien.ac.at

| | |
|---|---|
| Steuerberatungskanzlei Böck, Perchtoldsdorf | August 1989 bis Mai 1990<br>Buchhalterin |

Alexandra Mazak
Dipl.-Ing. Mag.
Lerchenfelderstraße 124-126/1/14 0664/75052744
A-1080 Wien mazak@ifs.tuwien.ac.at

## Bildungsweg

| | |
|---|---|
| Technische Universität, Wien | Sommersemester 2007 – Sommersemester 2012<br>Doktoratsstudium der technischen Wissenschaften<br>Dissertationsgebiet: Ontologien in der Informatik<br><br>Titel: CoMetO – a Cognitive Design Methodology for Enhancing the Alignment Potential of Ontologies.<br><br>Konzeption, Entwicklung, Implementierung (Prototyp) und Evaluierung (User Survey) einer kognitiv basierten Design Methodologie zur Umsetzung eines unidirektionalen, Hinweis basierten Kommunikationssystems vom Ontologie Entwickler zum User. Dabei wird die relationale Struktur einer Ontologie mit Kontext basierter kognitiver Semantik angereichert. Ziel ist die verbesserte Interpretation von Bedeutungsinhalten. Zusätzlich wurden Mismatch-at-Risk Metriken entwickelt die es dem User ermöglichen nicht lösbare Heterogenitäten (strukturelle, pragmatische Heterogenität) zwischen zwei zu alignenden Ontologien im Vorhinein abzuschätzen.<br>Nutzen: Kosten- und Zeitersparnis.<br><br>Forschungsschwerpunkte: Ontology Development, Ontology Re-Use (Ontology Alignment), Knowledge Acquisition, Cognitive-Engineering, Pragmatik, Relevanztheorie, Modelltheorie, Stochastik und Semantic Web. |
| Technische Universität, Wien | Wintersemester 2005/06 – Wintersemester 2006/07<br>Masterstudium Wirtschaftsingenieurwesen Informatik<br>Abschluss mit ausgezeichnetem Erfolg<br><br>Forschungsschwerpunkte: International Financial Reporting Standards (IFRS) mit dem International Accounting Standard - IAS 39, Financial Instruments: Recognition and Measurement, Delta-Hedging, Stochastik.<br><br>Schwerpunkte im Studium: Kosten- und Leistungsrechnung, Controlling (unter Unsicherheit) und Finanzrecht. Ingenieurswissenschaften - Bereiche aus Technischer, analytischer Chemie, Chemometrie (Statistik in der Chemie), Massenspektronomie und Verfahrenstechnik. |
| Technische Universität, Wien | Wintersemester 2005/06 – Wintersemester 2006/07<br>Masterstudium Informatikmanagement<br>Abschluss mit ausgezeichnetem Erfolg<br><br>Schwerpunkte im Studium: Kommunikation und sozial-wissenschaftliche Aspekte, Topisches Sozialsystem, systemisches Wissensmanagement, sowie Fachdidaktik der Informatik. |

Alexandra Mazak
Dipl.-Ing. Mag.
Lerchenfelderstraße 124-126/1/14 0664/75052744
A-1080 Wien mazak@ifs.tuwien.ac.at

| | |
|---|---|
| Technische Universität, Wien | Sommersemester 2002 – Sommersemester 2005<br>Bakkalaureat Data Engineering & Statistics<br><br>Schwerpunkte im Studium: Erfassen, analysieren und präsentieren von Daten aus unterschiedlichen Bereichen der Wirtschaft, Wissenschaft und Verwaltung, Statistik und Wahrscheinlichkeitsrechnung, Datenmodellierung, Datenbanken. |
| Technische Universität, Wien | Wintersemester 2001/02 – Wintersemester 2001/02<br>Universitätslehrgang Datentechnik. |
| Universität Wien, Juridicum | Wintersemester 1992/93 – Sommersemester 2000<br>Studium der Rechtswissenschaften |
| HBLA Biedermannsdorf | 1985 - 1989<br>5-jährige Ausbildung an der Höheren Bundeslehranstalt für wirtschaftliche Berufe. Abschluss Matura im Juni 1989. |