align++
A Heuristic-based Method for Approximating the Mismatch-at-Risk in Schema-based Ontology Alignment

Alexandra Mazak[1], Bernhard Schandl[3] and Monika Lanzenberger[1,2]
[1]*Institute of Software Technology and Interactive Systems, Vienna University of Technology, Vienna, Austria*
[2]*European Research Council, ERCEA, Covent Garden 21/23, Place Rogier, B-1049 Brussels, Belgium*
[3]*Department of Distributed and Multimedia Systems, University of Vienna, Vienna, Austria*

Keywords: Ontology alignment, Application context, Modeling focus, Heterogeneity coefficient, Mismatch-at-risk metric.

Abstract: Frequently, ontologies based on the same domain are similar but also have many differences, which are known as heterogeneity. The alignment of entities which are not meant to be used in the same context, or which follow different *modeling conventions*, may cause mismatch in ontology alignment. End-users would benefit from knowing the *risk level of mismatch* between ontologies prior to starting a time- and cost-intensive procedure. With our heuristic-based method *align++* we propose to consider the general application context of a modeled domain (the *modeling context*) in order to enhance the user support in schema-based alignment. In the method's first part, ontology concepts are enriched with weighting meta-information, resulting from two indicators: *importance weighting indicator* and *importance outdegree indicator*. These indicators contain *model-* and *graph-based* information and can be observed and measured at the schema level of an ontology. The output of the first part are ranking lists of *importance indicators* for each ontology concept in the role of a domain class. In the second part, the candidate sample for our *mismatch-risk model* bases on external user input by manually identifying concepts between the lists of each source ontology. The *heterogeneity risk* among the concepts' importance indicator values is measured as *standard deviation* over the candidate sample. Afterwards these measured values are aggregated, and a *heterogeneity coefficient* is calculated. On the basis of this risk factor the *mismatch-at-risk* (*MaR*) between ontologies can be approximated as a threshold for schema-based ontology alignment.

1 INTRODUCTION

An ontology is an artefact representing a scope of a real world domain for a specific purpose. In a collaborative modeling process multiple perspectives of a matter are condensed into a shared conceptualization. System analysts, in collaboration with domain experts, represent their view of the real world by using an abstract model, an ontology. Naturally, such models are marked by their authors' intentions and perspectives, and therefore cannot claim to represent objective reality. When a group of engineers start to conceptualize a certain domain they should agree on some shared representation forms, e.g., an expressive ontology language like OWL (Dean and Schreiber, 2004), and on a specific purpose for modeling this domain. This purpose (e.g., a certain business goal) restricts the modelers' views, and therefore the perspective on a domain. Ontology creators use entities to represent the domain of interest in a specific context, which results mainly from the purpose-specific usage of the domain. We call this specific context the *modeling context*. According to (Janiesch, 2010), "*when regarding modeling methods as social and contextualized complexes, it becomes necessary to include some stance of context in the meta model. [...], models or parts thereof can be equipped with context*".

Frequently, ontologies that describe the same domain of interest are similar but also expose many differences. These are known as *heterogeneity* and are rooted in diversity in ontology modeling. One reason for *conceptual* heterogeneity—which is also called *semantic* heterogeneity (Euzenat, 2001)—is the *difference in perspective* when modeling two ontologies (Euzenat and Shvaiko, 2007). Their example of maps addresses the problem of difference in per-

Mazak A., Schandl B. and Lanzenberger M. (2010).
align++ - A Heuristic-based Method for Approximating the Mismatch-at-Risk in Schema-based Ontology Alignment.
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 17-26
DOI: 10.5220/0003063600170026

spective from a spatio-temporal point of view. In (Benerecetti et al., 2001) the authors describe three kinds of perspectives: *spatio-temporal*, *logical*, and *cognitive*. Heterogeneity resulting from the first two kinds can be solved by DL-based techniques like SAT solver (Giunchiglia and Shvaiko, 2003). The *pragmatic* heterogeneity (Bouquet et al., 2004)—which is called *semiotic* heterogeneity by (Euzenat and Shvaiko, 2007)—results from differences in interpreting entities with regard to a specific context: "*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).

In our approach we focus on *semantics* from a cognitive perspective which leads to pragmatic heterogeneity problems in ontology alignment. Therefore, we prefer the notion *model-pragmatic* instead of *model-theoretic* semantics. The cognitive perspective includes the specific purpose of a modeled domain, and therefore it is related to the (intensional) *context layer* (Ehrig, 2007) of an ontology. Additionally, a possible mismatch risk can occur at the *ontology layer* which is called *explication mismatch* (Klein, 2001). This mismatch results from differences in *modeling conventions* (Chalupsky, 2000), which means dissimilarities in describing concepts. More detailed descriptions of heterogeneity and mismatch types have been given by (Visser et al., 1997), (Chalupsky, 2000), (Klein, 2001), and (Euzenat and Shvaiko, 2007).

Another problem in ontology alignment is to give end-users a quick and efficient overview of the source ontologies. Additionally, they should be supported to gain insight into the modeling process of those ontologies. A method which makes such an outline feasible can give users an idea about the application (modeling) context in which the entities are used for a specific purpose.

This paper is structured as follows: first, we describe the need of efficient aids for user support in schema-based ontology alignment. Then we introduce our heuristic-based method *align++* and present details about its two parts. We describe the idea of encoding context- and structure-based heterogeneity as possible *risk factors* in numerical values to approximate a *mismatch-at-risk* between ontologies. We finally underpin our research assumptions of align++ Part A with an evaluation survey.

2 APPROACH

In previous works we have proposed that in addition to the two factors *entity labels* and *relationships among entities* the *modeling focus* on entities should be additionally considered (Mazak et al., 2010). Analogous to the demand described by (Janiesch, 2010), "*[...] we attempt to systematize the current perceptions of context as relevant parameters for the adaption of conceptual modeling methods*"; and relating to (Ehrig et al., 2004), "*[...] similar entities are used in similar context*". In our approach the entities we focus on are the concepts of ontologies (or their classes, which are concrete representations of concepts, respectively). Our approach considers domain knowledge as meta-information in the form of two indicators, an *importance weighting indicator* and an *importance outdegree indicator* for classes. We denote with domain knowledge the modeling focus, which results from the context in which a certain domain has to be modeled.

Let us assume, for instance, that there are two ontologies (O_A and O_B) that describe the same domain of interest, a software tool for conference organization support (OAEI, 2009). We assume two different usage scenarios for these ontologies. In the first scenario, the purpose of creating both ontologies is to describe authors and their papers (Scenario 1). Therefore, the modeling focus of the ontology engineers is mainly on the concepts *Author*, *Contribution*, and *Article*, as well as these concepts' relations to other concepts. In the second scenario, the specific purpose of ontology O_A is to describe the events and organizations of the conference (Scenario 2), while the purpose of ontology O_B remains the same as in Scenario 1. Therefore, the modeling focus of ontology O_A in Scenario 2 is on the concepts *Working_Event*, *Administrative_Event*, and *Organization*. The context represents the environment in which the entities of an ontology have a certain level (*importance level*) of meaning. Thus, the introduced modeling context is equatable to the notion of *application context* (Ehrig et al., 2004). The differences due to the modeling focus cause semantic, pragmatic, and also terminological heterogeneity problems. Therefore, mismatch between ontologies may occur in the alignment process.

We have designed a heuristic-based method called *align++*, which follows the objective to support the end-user in ontology alignment by making heterogeneity between source ontologies visible before starting a schema-based alignment technique. The method provides a metric that quantifies the possible mismatch between ontologies. It helps users to gain a better understanding of ontologies, and disburdens them from complex, time-, and cost-intensive tasks. The name align++ results from the two steps in which this method is divided, an *ex ante* and an *ex post* step.

Firstly, using the techniques of the ex ante step of

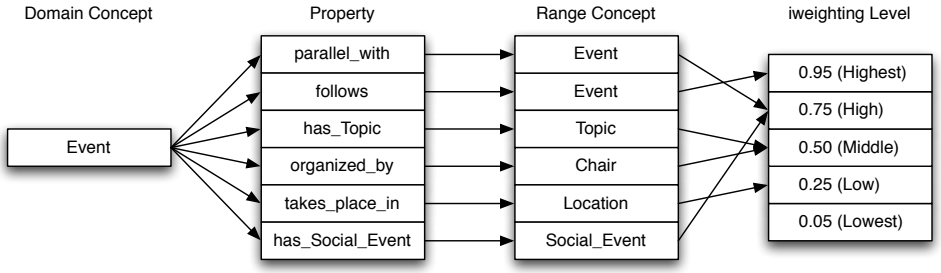

Figure 1: Example recording the importance-weighted `owl:Class` *Event* at the schema level (TBox) of an ontology.

Part A, information that results from the context and ontology layer of an ontology can be observed and measured. Secondly, each domain concept is annotated with these measurements in the form of meta-information by weighted values. The concepts with their labels and computed values are recorded as ordered *ranking lists*. Possible heterogeneity factors resulting from the individual process of meta-modeling ontologies at the schema level are mapped to their concepts. Thus, we enrich the element level with meta-information of the structure level.

The ex post step of *Part B* starts with a user selection of similar concepts out of the ranking lists of two or more ontologies as input for our *mismatch-risk model*. This strategy of a manually conducted concept selection minimizes a possible structural falsification induced by other methods, e.g., lexical matching techniques. After this user selection we evaluate the *heterogeneity risk* measured as *standard deviation* among the concepts' importance indicator values, aggregate the measured values, and calculate a *heterogeneity coefficient*. On the basis of this risk factor the *mismatch-at-risk* (*MaR*) between the source ontologies can be approximated as a threshold value for the schema-based alignment process.

2.1 Part A: Evaluating Risk-determining Indicators

We use OWL DL (Dean and Schreiber, 2004) as vocabulary to describe domains of interests. There, an ontology is a set of *logical axioms* that are asserted in the *TBox* at the schema level. With our method we focus on the more general context of these logical axioms or statements, rather than on situational details of the *ABox* at the *instance level*. We agree with (Janiesch, 2010) in that the use of *situational context* is too detailed to allow for a meaningful reuse in ontology alignment. Therefore, our method considers only schema level information but no instance data. We further assume that the modeling context is mainly hidden in the relational structure of an ontology (cf. Section 3). According to (Euzenat and Shvaiko, 2007), "*matching ontologies from their relational (or external) structure is very powerful [...]*", and "*it is worth considering what are the important relations before using such techniques*"—meaning techniques which consider the relational structure of an ontology.

The modeling focus is not directly observable and measurable, hence we need indicators that quantify the level of meaning encoded in these schemas for further computation. For this purpose we introduce two indicators: the *importance weighting indicator* (IwI_c) and the *importance outdegree indicator* (IoI_c) of ontology concepts (c). As introduced in our previous works (Mazak et al., 2010), $IwI_c \in [0;1]$ results from the importance-weighted (model-pragmatic) semantics of binary relations (`owl:ObjectProperties`) depending on their particular domain/range combinations. According to (Euzenat and Shvaiko, 2007), *"the semantics of ontologies can be constrained by additional axioms"*, which are in our case the `rdfs:domain` and `rdfs:range` assertions that constrain an `owl:ObjectProperty` (Horridge, 2004). This means that the local semantics (meaning) of a statement is constrained based on its purpose-specific usage. This information is mainly encoded in the relations between concepts (`owl:ObjectProperties`) and not only in their taxonomic relations.

The first step of the weighting procedure manually conducted by the ontology engineers during the ontology design and development process, since "*semantics is usually specified explicitly at design time*" (Shvaiko and Euzenat, 2004). As an aid for setting importance weighting levels in this procedure the ontology engineers could be geared to the *competency questions* in (Grüninger and Fox, 1995) and (Noy and McGuinness, 2001). The importance weighting procedure is practicable for the ontology developers also in large ontologies. Since an importance weight is annotated, with a simple point-and-click interaction (Mazak et al., 2010), when the object property with its domain/range constraints is generated at design time. IwI_c values encode the usage of entities in a certain application context. This context layer meta-information is annotated on the *relation signature* σ_R :

Table 1: Importance Weighted Indicator (IwI_c) values for ontologies O_A and O_B.

(a) Scenario 1: equal modeling focus.

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

(b) Scenario 2: different modeling focus.

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

$R \rightarrow C \times C$ (Ehrig et al., 2004) at the schema level. Therefore, the level of context-based heterogeneity—being a possible risk factor in the alignment process—is encoded in the value of a concept's IwI_c.

In a second step we identify the importance outdegree indicator values (IoI_c, $IoI_c \in [0;1]$), which result from a weighting based on the outdegree of a concept c in proportion to the concept with the highest outdegree within the ontology. Therefore, this second indicator considers a possible heterogeneity risk resulting from differences in describing concepts. More precisely, in the values of IoI_c the heterogeneity risk of a concept based on differences in *modeling conventions* (Chalupsky, 2000) is encoded. Additionally, IoI_c indicates the importance of concepts for structure-based alignment techniques (e.g., graph-based methods). Such information is important for users to detect efficient initial points for starting alignment or mapping methods like Anchor-PROMPT (Noy and Musen, 2001).

Figure 1 shows an excerpt of an ontology that has been enriched with IwI_c indicators for instance, the relation: *"Event follows Event"* has been weighted with highest importance, while the relation: *"Event takes_place_in Location"* is only of low importance. As can be seen from this figure, the indicator-based values of concept relations can be stored in multistage hash maps. In the current version of align++—which is implemented using the Eclipse environment (Gronback, 2009)—, concepts with their weighted values and labels are recorded in form of ordered lists, which can be additionally used as *ranking lists* (cf. Table 1 and Table 2).

2.2 Part B: Exploiting Risk Factors during Ontology Alignment

The second part of align++, the ex post step, is initiated at the beginning of an alignment between two ontologies. To describe this part in detail we start with an example on the basis of the ontologies *confOf* (O_A) and *crs_dr* (O_B), and the two scenarios we have described in Section 2. Further, we assume that all logical statements of these two ontologies have already been importance-weighted (cf. Section 2.1) based on the respective scenario, and that for each domain concept the IwI_c and IoI_c-based values have been computed.

Before they start an alignment process, end-users should be supported so that they can get a quick and context-based overview of the source ontologies. They should be able to easily detect the core concepts of these ontologies, their importance in a certain application context, and whether concepts are efficient candidates to be selected as initial points for graph-based alignment methods or propositional techniques (e.g. SAT solver).

Table 1 shows the lists in which the concepts of O_A and O_B are ranked by their indicator-based values. The ranking bases on an average of the values resulting from the weighting procedures, which have been manually conducted by the survey respondents (cf. Section 3). On the one hand, end-users can easily detect the core concepts *Author* and *author*, which are also syntactically similar, and *Contribution* and *article* (Table 1a). As we can see, these lists 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* and *article*. On the other hand the list depicted in Table 1b shows dif-

Table 2: Importance Outdegree Indicator (IoI_c).

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

ferences in the ranking of concepts. These differences are identifiable due to the differences of their IwI_c-based values. It is evident that both ontologies describe the same domain of interest, but with a different modeling focus on it. Moreover, the user can detect that the intended usage of the concepts may differ. Thus, they can derive that the application context is probably not the same (i.e., *pragmatic heterogeneity*). We assume that this kind of heterogeneity mainly results from the interpretation of entities by humans due to a certain application context: "*this kind of heterogeneity is difficult for the computer to detect and even more difficult to solve, because it is out of its reach*" (Euzenat and Shvaiko, 2007).

Figure 2 presents the differences in the structures of our two example ontologies, which are reflected in their respective IoI_c values. O_A consists of a large number of classes, which are arranged in three hierarchy levels. O_B has significantly fewer classes with only two levels of hierarchy. This kind of *structural* heterogeneity results from differences in describing concepts: "*[...] a distinction between two classes can be modeled using a qualifying attribute or by introducing a separate class*" (Klein, 2001). (Chalupsky, 2000) denotes this kind of heterogeneity as differences in *modeling conventions*. Table 2 presents this possible heterogeneity factor in form of ranked concepts based on their IoI_c-based values. These values encode graph-based information; in particular, the outdegree of a concept in proportion to the concept with the highest outdegree in each of the source ontologies. In our example, the concepts *Person* of O_A and *article* of O_B have the most outgoing relations to other concepts.

All these kinds of differences or heterogeneity cause mismatch between ontologies in their alignment. It would be a benefit for end-users to know the *risk level* of mismatch (or *mismatch-at-risk*) *MaR* before starting a time- and cost-intensive schema-based alignment process: "*in real-world applications, schemas/ontologies usually have both well defined and obscure labels (terms), and contexts they occur, therefore, solutions from both problems would be mutually beneficial*" (Shvaiko and Euzenat, 2004). Therefore, we introduce a statistical method in this second part of align++ which we call *mismatch-risk model*. Using this method, a possible mismatch-at-risk (*MaR*) between source ontologies, which results from heterogeneity factors at the context and ontology layer, can be approximated.

(Shvaiko and Euzenat, 2004) point out that semantics is usually given in a structure and not at the element level. The first input of our risk model exploits (local) model-based semantic and graph-based syntactic meta-information from the structure level of ontologies annotated on their concepts at the element level. This *internal* input results from the semi-automated computations of the IwI_c and IoI_c in Part A of our method (cf. Section 2.1). According to (Euzenat and Valtchev, 2004), "*to provide the most complete basis for comparison, one may wish to bring knowledge encoded in relation types to the object level*". Therefore, align++ considers structure level meta-information encoded at the element level to approximate the *MaR* as an efficient benchmark. Additionally, according to the *process dimension* described by (Shvaiko and Euzenat, 2004) the input is interpreted by an *external resource* in the form of human input. Therefore, the input for the mismatch-risk model is both internal and external.

The risk model needs *external* user input for

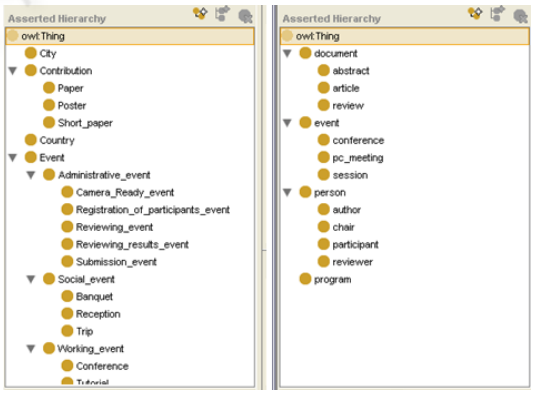

Figure 2: Two differently structured ontologies describing the same domain of interest, visualized in Protégé.

manually identifying matching candidates from the ranking lists of each source ontology. This strategy of a manually conducted concept selection minimizes the risk of information loss, resulting from possibly poor quality produced by automated methods. For instance, in Scenario 1 (Table 1a) a user can easily detect the correspondence between the concepts *Contribution* of O_A and *article* of O_B, whereas lexical

matching methods cannot accomplish this.

According to (Euzenat and Shvaiko, 2007) the technique of manually determining the candidate sample can be classified under *repository of structures*. One approach of this technique has been introduced by (Rahm et al., 2004). This *fragment-oriented* approach decomposes a large matching problem into smaller sub-problems on *schema fragments*, based on a divide-and-conquer strategy. Therefore, schema elements become special schema fragments. Various types of schema information will be exploited by this approach, as well as background knowledge. The purpose of this *fragment-oriented* approach for our method is to determine an efficient candidate sample for the mismatch-risk model as input.

It can be assumed that, while experienced ontology engineers will expect a certain level of heterogeneity between ontologies, they have no means to validate their expectations before they actually start the alignment process, leading to uncertainty or the risk of mismatch between source ontologies.

To address this, we adopt the *value at risk* (*VaR*) metric, which is a widely used risk measure in financial mathematics (Franke et al., 2004), for Part B of our align++ method. We analyze the *variation* of the indicator-based values among the concepts of the candidate sample to approximate the mismatch-at-risk (*MaR*) between the source ontologies. As the variation among these values increases, the probability for *MaR* grows. Therefore, the risk factor in our method is the *margin of deviation* among the indicator-based values. This margin of deviation (the *heterogeneity risk*) is similar to the *volatility risk* in financial markets. This heterogeneity risk can be denoted as a *(continuous) random variable* (X). A widely used risk measure that provides a quantified estimate of uncertainty is the standard deviation $\sigma(X)$, defined as

$$\sigma(X) = \sqrt{\mathrm{E}[(X-\mu)^2]} \tag{1}$$

More formally, let X be a random variable with mean value $\mathrm{E}(X) = \mu$. The operator E denotes the expected value of X. The standard deviation (σ) is the square root of the expected value of $(X-\mu)^2$ (Stahel, 2000). In a further step we aggregate the measured values to compute the *median absolute deviation* as a robust estimator of variation. We call this median, which is a reliable measure of uncertainty, the *heterogeneity coefficient* of the sample. On the basis of this coefficient as a cumulated risk factor we can approximated the *MaR*.

In financial mathematics, the *VaR* risk metric summarizes the distribution of possible losses by a *quantile*, i.e., a point with a specified probability of higher losses (Franke et al., 2004). To adopt this approach we assume that our candidate sample underlies a normal distribution. Thus, we convert the random variable X with its parameters μ and σ to a random variable Z with expectation $\mathrm{E}(Z) = \mu = 0$ and $\sigma = 1$, using a transformation to standardize X (Meintrup and Schäffler, 2005):

$$Z = \frac{X-\mu}{\sigma} \tag{2}$$

The risk positions in the mismatch-risk model are the IwI_c-based values of the candidates (concepts) in the sample determined by the user, or the IoI_c-based concept values, respectively. The margin of deviation or variation among these values are the realizations of Z. The output of the mismatch-risk model is the approximated *MaR* between all (i.e., two or more) ontologies. This is a threshold value such that the probability that the variation gets "unfavorable" because it exceeds this value, is the given value. The *MaR* between source ontologies can be calculated by *quantiles*, similar to the value-at-risk metric in financial statistics (Eller et al., 2002). The normalization (cf. Equation 2) helps us to calculate these quantiles, since it is easier to determine the *scaling factor* for a certain confidence level (e.g, 95% or 99%) by the quantile of the standard normal distributed random variable Z.

In our example, let us assume the user detects the correspondences between the concepts *Author/author* and *Contribution/article* and selects these concepts as the candidate sample for the mismatch-risk model. On the basis of this *external* user input we calculate a heterogeneity coefficient as the median based on the variation among the IwI_c-based values of these candidates. Table 3a presents this measure of uncertainty for Scenario 1, where the modeling focus is equal for both ontologies. The heterogeneity coefficient is 0.04, which indicates a very low risk factor. In contrast, the coefficient of Scenario 2 (cf. Table 3b), where the modeling focus is different, indicates a high risk factor. On the basis of these heterogeneity coefficients we calculate the *MaR* for both scenarios for a 95%-confidence level. The 95%-quantile of a standard normal distributed random variable Z lies in a defined range between 1.64 and 1.65. Therefore, the scaling factor for the calculation of the *MaR* in both tables is 1.64. The *MaR* for Scenario 1 (Table 3a) approximates a very low threshold value, while in Scenario 2 (Table 3b) the mismatch-at-risk is highest with 88%. Thus, it would be a better choice for the user to align the ontologies in Scenario 1 than in Scenario 2.

In analogy to the calculation of the mismatch-at-risk based on the heterogeneity at the context layer between these ontologies, the heterogeneity risk at the ontology layer resulting from the variation of the IoI_c-based concept values can be computed.

Table 3: Calculation of Heterogeneity coefficients and Mismatch-at-risk levels for both scenarios.

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

IwI of concept O_A		O_B		Standard deviation of IwI_c-based values
Author	0.95	author	0.92	0.02
Contribution	0.92	article	0.83	0.06
Heterogeneity coefficient				0.04
MaR with 95% confidence level				**7%**

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

IwI of concept O_A		O_B		Standard deviation of IwI_c-based values
Author	0.13	author	0.92	0.56
Contribution	0.11	article	0.83	0.51
Heterogeneity coefficient				0.53
MaR with 95% confidence level				**88%**

3 EVALUATION

We conducted an evaluation of align++ by a questionnaire-survey which we mailed to 20 respondents. 5 female and 13 male respondents completed the questionnaires, which were anonymized for the analysis process. 12 of these 18 participants were researchers in Computer Science, while 4 respondents were students in the fields of Computational Intelligence, Software & Information Engineering, and Information & Knowledge Management. Further, two respondents were employees in leading positions at a software house. 12 respondents declared themselves to be well-versed in ontology engineering and alignment, while the others declared themselves as versed.

In the course of this survey the respondents were asked to weight all logical statements (`owl:ObjectProperties` with their domain and range axioms) of the two example ontologies *confOf* (O_A) and crsdr (O_B). For this purpose a simple point-and-click user interface was implemented. With the aid of our assumed scenarios (cf. Section 2.2) we have predefined the respective application (modeling) context. In order for respondents to know how to weight each axiom with an importance level a brief introduction with examples was included with the survey questionnaire.

The very low variation among the IwI_c-based values of concepts in Scenario 1 emphasizes our assumptions made in Section 2, whereas the very high variation of those values among the same concepts in Scenario 2 reflects the heterogeneity and mismatch problems that were described in Section 1 of this paper.

A further result of the participants' weighting procedures is that all of the 18 respondents have weighted the axioms in a nearly equal manner due to the given modeling focus, which can be seen from Table 4. In Scenario 1 the concepts *Author* and *Contribution* of O_A as well as *author* and *article* of O_B have nearly equal IwI_c-based mean values, represented in Table 3a. If the predefined focus is on authors and their papers (Scenario 1) the relations where these concepts participate in the role of a domain class are weighted highest, which results in calculated IwI_c-based means with values of 0.83 and 0.95. Otherwise, if the focus of O_A was on events and organizations (Scenario 2) the binary relations in which the concepts *Author, Contribution* participate are weighted lowest, which results in IwI_c-based mean values of 0.13 for *Author* and 0.11 for *Contribution*. From this it follows that the standard deviation between the IwI_c-based values of the concepts *Author/author* and *Contribution/article* is lowest with values of 0.02 and 0.05 in Scenario 1 and highest in Scenario 2. Table 4 represents the equalities and differences of the weighting annotations per respondent and points out that the application context (modeling context) restricts a modeler's view (cf. Section 1 and Section 2).

Table 4: Importance Weighting Indicator (IwI_c), calculated from the 18 respondents' property weightings.

	Ontology O_A		Ontology O_B			
	Both Scenarios		Scenario 1		Scenario 2	
Respondent	*author*	*article*	*Author*	*Contribution*	*Author*	*Contribution*
1	0.95	0.84	0.95	0.95	0.05	0.05
2	0.95	0.90	0.95	0.95	0.25	0.15
3	0.95	0.79	0.95	0.95	0.05	0.15
4	0.95	0.90	0.95	0.95	0.05	0.15
5	0.95	0.79	0.95	0.85	0.25	0.05
6	0.95	0.79	0.95	0.85	0.05	0.15
7	0.95	0.84	0.85	0.85	0.25	0.05
8	0.95	0.84	0.95	0.85	0.25	0.15
9	0.95	0.78	0.95	0.85	0.05	0.15
10	0.95	0.85	0.85	0.85	0.05	0.05
11	0.95	0.84	0.85	0.95	0.25	0.15
12	0.95	0.79	0.95	0.95	0.25	0.15
13	0.95	0.79	0.95	0.95	0.05	0.05
14	0.95	0.85	0.95	0.95	0.05	0.05
15	0.95	0.85	0.95	0.95	0.05	0.15
16	0.95	0.79	0.95	0.95	0.25	0.05
17	0.95	0.85	0.85	0.95	0.05	0.05
18	0.95	0.85	0.85	0.95	0.05	0.15
mean	0.92	0.83	0.95	0.92	0.13	0.11

After the weighting procedure the participants were asked further questions. They were asked to answer them on a 5-level Likert scale (*strongly disagree*, *disagree*, *undecided*, *agree*, *strongly agree*). In the

following we present an overview on the ratings and explanatory statements given by the 18 respondents.

89% strongly agree that the modeling focus of an ontology and its entities depends on a certain perspective ontology engineers have in mind when conceptualizing a domain of interest. They further state that due to semantic relativism, as already known in database engineering, models are always subjective, which causes heterogeneity problems in the alignment of these models. 75% strongly agree, and 17% agree that the meaning of ontology concepts and their context-sensitive (purpose-specific) usage mainly depends on this modeling focus. Additionally, they agree that the common understanding of engineers which bases on the application of the ontology is important. One of the participants mentions, *"it is not possible to model anything without the influence of context-sensitive parameters"*. Another respondent states, *"a concept can be very important in one relation, and unimportant in another depending on the modelers' foci"* (cf. Table 3). This feedback corresponds with our assumptions as described before.

Answering the question whether there are other components on which the meaning of concepts depends the majority of the respondents reply with "yes". According to the participants these components include *"experiences, culture, stakeholders, background of engineers, skills, environmental parameters, intended audience"*. Therefore, we have to state more precisely that with the expression "the modeling focus mainly depends on a certain modeling context" we denote the application context (Ehrig et al., 2004) of an ontology as mentioned in Section 2. This is the context in which an ontology and its entities are modeled for a purpose-specific usage (e.g., a certain business goal).

The participants were further asked whether they agree that the logical statements among concepts are an indicator for their context-sensitive usage. 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 states that also the taxonomic structure, which is commonly used in ontology alignment should be considered, too. We assume in our approach that the *local context* of concepts (i.e., their outgoing relations—`owl:ObjectProperties`—to other concepts within the ontology) is more important (cf. Section 2.1).

All respondents strongly agree that for instance, the importance weighting degree for the relation *Author* $\rightarrow$ *writes* $\rightarrow$ *Contribution* would be different if the ontology engineers' modeling focus was on the authors rather than on the conference program. This fact additionally points out that semantic as well as semiotic heterogeneity can in fact be made visible by our approach, as described in detail in Section 2.2.

In our approach, ontology modelers can choose between five degrees of weighting labels: *Highest*, *High*, *Middle*, *Low*, and *Lowest Importance*. We think that users prefer to assign importance labels instead of numerical values. 13 of the respondents state that five degrees are enough, 4 consider three as sufficient, and 1 respondent indicates that a finer-grained schema would be better. We think that five degrees, including a neutral level, are a reasonable compromise to convey an importance weighting to the user. However, since these five levels are mapped to the continuous interval $[0;1]$ the approach allows one to arbitrarily increase or reduce the number of degrees.

All participants strongly agree that the concepts' ranking lists, which base on the introduced indicators, are efficient to give end-users a quick and context-based overview about the core concepts of the source ontologies. Further, they strongly agree that due to the indicator-based concept values end-users are able to easily detect possible differences in the application context or modeling focus, respectively (cf. Section 2.2). The majority of the respondents point out that it may be useless to align ontologies with different perspectives on their entities. Finally, they strongly agree that the IwI_c- and IoI_c-based concept values are efficient indicators for possible heterogeneity risks between source ontologies in schema-based ontology alignment.

4 RELATED WORK

Detailed surveys about techniques which also use weights in their approach have been given by (Euzenat and Shvaiko, 2007) and by (Ehrig, 2007). Some of those techniques consider solely *is-a* relationships among concepts, while others (e.g., statistical methods) exploit the instance data of ontologies. These instances serve as representative samples to take measurements on which comparisons between two source ontologies can be established. In our approach we advance a view, corresponding with (Janiesch, 2010), that the situational context at the instance level is too detailed to allow a meaningful reuse; therefore, we consider only schema level information.

The semantics of an *is-a* taxonomy is exploitable by counting the paths within the hierarchy. The weighting of such a taxonomy is mainly computed by fixed values (e.g., 0.5, or 1) for each path length, depending on the distance from the root. A consideration of object properties themselves is not useful to

make meaningful statements.

In our approach we consider `owl:ObjectProperties` with their domain and range axioms to make use of their semantics. Automatic ranking methods (Wu et al., 2008) identify the importance of concepts by counting the number of relations of one concept to another in a first step, while also taking into account the other concept's importance. However, a method that aims to consider the concept importance in a certain application context requires non-trivial knowledge about the modeled domain. Thus, our method already starts its weighting procedure (cf. Section 2.1) during the ontology design and development process. In our opinion nobody is better qualified to annotate ontologies with weighting factors than ontology engineers themselves. Another benefit of the approach is that the manually annotated weighting labels are specific values for each logical statement, instead of fixed values as in other methods.

Semantic-based techniques often build on intermediate formal ontologies to define a common context or background knowledge in order to bridge the gap caused by the lack of a common ground for comparison. This common ground can often be found in external resources and models (e.g., DOLCE, WordNet). These methods help in handling the disambiguation of multiple possible meanings of terms. In the align++ method such *oracles* are not required. We involve the end-users as an *external resource* to detect similar concepts on the basis of the ranking lists output by Part A. These lists help end-users to define efficient candidate samples for the mismatch-risk model. The consideration of different heterogeneity types between source ontologies as possible risk factors, before starting an alignment, is new also the approximation of a mismatch-at-risk between ontologies in schema-based alignment.

5 CONCLUSIONS AND FUTURE WORK

The approach we provide is a heuristic-based method to make heterogeneity visible for end-users before starting time- and cost-intensive schema-based alignment methods. With our method the risk level of a possible mismatch between ontologies can be approximated in the form of a mismatch-at-risk (*MaR*) value. Therefore, if two or more ontologies are available for alignment the user can choose those two ontologies with a minimum *MaR*. Otherwise, if only two ontologies of a certain domain are existing the benefit for the user is to know about the mismatch risk before aligning them. Additionally, our presented method supports users in a better understanding of the source ontologies by providing a quick and context-based overview of these ontologies by ranking lists of their concepts. Currently, we conduct a detailed user evaluation of align++ Part B; for the future, we aim to extend our approach by considering more elements of the respective ontologies (e.g., taxonomy relationships) in the calculation of the heterogeneity coefficient.

ACKNOWLEDGEMENTS

Our special thanks go to *Secure Business Austria Research GmbH* for their pecuniary aid in the course of the FAMOS-Project (Female Academy for Mentoring, Opportunities and Self-Development).

REFERENCES

Benerecetti, M., Bouquet, P., and Ghidini, C. (2001). On the dimensions of context dependence. In *Third International and Interdisciplinary Conference, CONTEXT*, Dundee (UK).

Bouquet, P., Euzenat, J., Franconi, E., Serafini, L., Stamou, G., and Tessaris, S. (2004). *D2.2.1: Specification of a common framework for characterizing alignment*. Knowledge Web Consortium.

Chalupsky, H. (2000). OntoMorph: A translation system for symbolic logic. In Anthony G. Cohn, F. G. and Selman, B., editors, *KR2000: Principles of Knowledge Representation and Reasoning*, pages 471–182, San Francisco, CA.

Dean, M. and Schreiber, G. (2004). *OWL Web Ontology Language Reference (W3C Recommendation 10 February 2004)*. World Wide Web Consortium.

Ehrig, M. (2007). *Ontology Alignment: Bridging the Semantic Gap*, volume 4 of *Semantic Web And Beyond Computing for Human Experience*. Springer, 1st edition.

Ehrig, M., Haase, P., Hefke, M., and Stojanovic, N. (2004). Similarity for Ontologies - a Comprehensive Framework. In *In Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, at PAKM 2004*, Vienna (Austria).

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

Euzenat, J. (2001). Towards a Principled Approach to Semantic Interoperability. In *Workshop on Ontologies and Information Sharing, IJCAI01*, Seattle (WA US). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.9779.

Euzenat, J. and Shvaiko, P. (2007). *Ontology Matching*. Springer, Heidelberg (DE).

Euzenat, J. and Valtchev, P. (2004). Similarity-based ontology alignment in OWL-Lite. In *The 16th European Conference on Artificial Intelligence, ECAI-04*, Valencia (Spain).

Franke, J., Härdle, W., and Hafner, C. (2004). *Einführung in die Statistik der Finanzmärkte*, volume 2 of *Statistik und ihre Anwendungen*. Springer, 1st edition.

Giunchiglia, F. and Shvaiko, P. (2003). SEMANTIC MATCHING. Technical Report DIT-03-013, University of Trento Department of Information and Communication Thechnology, 38050 Povo, Trento (IT), Via Sommarive 14. http://eprints.biblio.unitn.it/archive/00000381/01/013.pdf.

Gronback, R. C. (2009). *Eclipse Modeling Project: A Domain-specific Language Toolkit*. Addison-Wesley, 1st edition.

Grüninger, M. and Fox, M. S. (1995). Methodology for the Design and Evaluation of Ontologies. In *International Joint Conference on Artificial Inteligence IJCAI95, Workshop on Basic Ontological Issues in Knowledge Sharing*, Toronto (CA).

Horridge, M. (2004). *A Practical Guide To Building OWL Ontologies With The Protege-OWL Plugin*. University of Manchester, 1 edition. http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/.

Janiesch, C. (2010). 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.

Klein, M. (2001). Combining and Relating Ontologies: An Analysis of Problems and Solutions. In Gomez-Perez, A., Gruninger, M., Stuckenschmidt, H., and Uschold, M., editors, *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle (WA).

Mazak, A., Schandl, B., and Lanzenberger, M. (2010). Enhancing Structure-based Ontology Alignment by Enriching Models with Importance Weightings. In *3rd International Workshop on Ontology Alignment and Visualization (OnAV'10)*, Krakow (Poland).

Meintrup, D. and Schäffler, S. (2005). *Stochastik*. Statistik und ihre Anwendungen. Springer, 1st edition.

Noy, N. F. and McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Technical Report SMI-2001-0880, Stanford University, Stanford (CA), 94305.

Noy, N. F. and Musen, M. A. (2001). 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).

OAEI (2009). OAEI-2009 Campaign Conference track. http://oaei.ontologymatching.org/2009/conference/.

Rahm, E., Do, H.-H., and Maßmann, S. (2004). Matching Large XML Schemas. In *SIGMOD Record*, volume 33. ACM.

Shvaiko, P. and Euzenat, J. (2004). A Survey of Schema-based Matching Approaches. Technical Report DIT-04-087, University of Trento, Department of Information and Communication Technology.

Stahel, W. A. (2000). *Statistische Datenanalyse*. Vieweg & Sohn Verlagsgesellschaft mbH, 3rd edition.

Visser, P. R. S., Jones, D. M., Bench-Capon, T., and Shave, M. (1997). An analysis of Ontology Mismatches; Heterogeneity versus Interoperability. In *AAAI 1997, Spring Symposium on Ontological Engineering*, Stanford (CA US). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.6709.

Wu, G., Li, J., Feng, L., and Wang, K. (2008). Identifying Potentially Important Concepts and Relations in an Ontology. In *Proceedings of the 7th International Conference on The Semantic Web*, Karlsruhe (Germany).