Let's Agree to Disagree: On the Evaluation of Vocabulary Alignment

Anna Tordai, Jacco van Ossenbruggen, Guus Schreiber, Bob Wielinga
Department of Computer Science
VU University Amsterdam
De Boelelaan 1081a, 1081 HV
Amsterdam, Netherlands
{a.tordai, j.r.van.ossenbruggen, guus.schreiber, b.j.wielinga}@vu.nl

ABSTRACT

Gold standard mappings created by experts are at the core of alignment evaluation. At the same time, the process of manual evaluation is rarely discussed. While the practice of having multiple raters evaluate results is accepted, their level of agreement is often not measured. In this paper we describe three experiments in manual evaluation and study the way different raters evaluate mappings. We used alignments generated using different techniques and between vocabularies of different type. In each experiment, five raters evaluated alignments and talked through their decisions using the think aloud method. In all three experiments we found that inter-rater agreement was low and analyzed our data to find the reasons for it. Our analysis shows which variables can be controlled to affect the level of agreement including the mapping relations, the evaluation guidelines and the background of the raters. On the other hand, differences in the perception of raters, and the complexity of the relations between often ill-defined natural language concepts remain inherent sources of disagreement. Our results indicate that the manual evaluation of ontology alignments is by no means an easy task and that the ontology alignment community should be careful in the construction and use of reference alignments.

Categories and Subject Descriptors: I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods [semantic networks]

General Terms: Experimentation

Keywords: empirical study, inter-rater agreement, manual evaluation, vocabulary alignment

1. INTRODUCTION

In this paper we study the quality of manual evaluation of ontology alignments. Manual evaluation is a fundamental method for establishing quality in ontology and vocabulary alignment and many other fields such as information retrieval and linguistic research. In vocabulary matching, evaluators rate the quality of mappings by assigning them into categories, thus creating a gold standard, also called a reference alignment that is used to measure the quality of mapping algorithms. An established method of validating the gold standard is to have multiple raters evaluate the same set of mappings into categories. Agreement between raters is then measured by correcting for chance agreement using measures such as Cohen's kappa [6]. Given a high enough inter-rater agreement measure the results of the manual evaluation can be used as a gold standard. However, what the threshold of agreement should be is not clear cut and also depends on the research field in question [12, 4, 2].

While evaluation by multiple raters is a preferred validation method, it is not always documented in practice. The focus of evaluation reports frequently lies on the performance of evaluated tools. In cases where inter-rater agreement measures have been used in the manual evaluation, the reported levels of agreement diverge greatly. For example, in the Very Large Cross-Vocabulary track of the Ontology Alignment Evaluation Initiative (OAEI)[1] organizers reported perfect agreement between raters [8]. Halpin et al. [9] however reported very poor agreement levels in their experiments evaluating `owl:sameAs` mappings sampled from Linked Data. In our previous work [15], and [16] we also measured interrater agreement and found only moderate levels of agreement between raters which we found unexpectedly low. As manual evaluation is such an integral part of the evaluation process we have asked ourselves why raters find it so difficult to agree on relationships between concepts. In this paper we will focus on the following research questions:

1. What is the level of agreement between raters when evaluating alignments?

2. If agreement is low, what are the reasons behind it?

To this end we performed three evaluation experiments on mappings between two sets of vocabularies and analyzed the results quantitatively as well as qualitatively. Because our experiments were explorative in nature, we only evaluated small sets of mappings and focused on qualitative analysis in particular. As part of our experimental setup we created specific guidelines detailing evaluation categories and provided examples and further explanations to raters. We performed a quantitative analysis by using established measures such as Cohen's kappa and Krippendorff's alpha and analyzed data from "think aloud" sessions during the experiments.

2. RELATED WORK

There are relatively few research papers on vocabulary alignment that detail an evaluation by multiple raters and include inter-rater

[1] `http://oaei.ontologymatching.org/`

K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.

agreement measurements. In previous papers [15, 16] we described case studies of alignments between various vocabularies, including WordNet, and manually evaluated resulting mappings. In both case studies we validated our evaluation by asking three raters to evaluate samples of the alignments. We used an evaluation tool to display mapped concepts along with their immediate hierarchies, scope-notes and labels, and had raters select a SKOS matching relation [13] to categorize the mapping. To support raters in this task we provided a set of guidelines which included a short description of each matching relation based on the W3C recommendation and examples of mappings. We measured Cohen's kappa and found moderate agreement (0.56) between raters in our first report[15], and just slight agreement (0.36) in [16]. As our goal in both case studies was to assess precision with regards to equivalence, we reduced the number of categories into equivalent or not equivalent. With just two categories the inter-rater agreement rose to substantial agreement (0.70 and 0.67). From these values we concluded that the evaluation task is difficult even for humans, and further study revealed that raters' understanding of SKOS matching relations varied from person to person. We also found that the lexical richness of vocabularies like WordNet may contribute to the difficulty level of the evaluation task, as closely related senses are separated into different concepts. However, a clearer delineation between mapping relations would likely raise agreement.

Halpin et al. [9] also reported low levels of agreement in their paper. They analyzed the use of `owl:sameAs` mappings in Linked Data and defined a similarity ontology to differentiate between various degrees of similarity. In their evaluation experiment they defined 5 levels of similarity relations between entities and used them to evaluate mappings. The agreement level between raters was very low with kappa at 0.16 and the authors attributed this to different styles of judgments. After a recombination of the rating categories into three the agreement increased to 0.32, which is still lower than what we experienced. They found that raters had the most difficulty in defining whether two entities were the same, and that background knowledge has an impact on decisions. They concluded that this inability to rate entities as the same stemmed from not knowing how the entities would be used. In the mapping categorization instruction Halpin et al. used variations on the same type of entity to illustrate each mapping category (descriptions of performances of Bohemian Rhapsody by Queen or some other band). In richly varied data such as Linked Data mapping categories need to be defined in more general terms with examples varying in domain and type. Raters then have less need to interpret examples themselves.

While guidelines with clearer descriptions of categories could improve inter-rater agreement, it is clear from these reports that the task of manual evaluation is difficult. Manual evaluation of mappings is a type of categorization task. Studies in cognitive science in general [11] and linguistic categorization in particular[14] have shown that humans do not categorize according to classic Aristotelian view where each category is clearly defined and mutually exclusive. Instead, Lakoff argues in his book [11] that prototype theory is at the core of cognitive categorization whereby some members of a category are more central (prototypical) than others. For example, *chair* is a more prototypical member of the category *furniture* than a *side-table*. Categories thus form a graded cloud with fuzzy boundaries where member concepts do not necessarily share common properties. They are defined by culture and experience and therefore vary from person to person. This fuzzy nature of categories provides an insight into why categorization tasks can be difficult.

3. EXPERIMENTAL SETUP, TOOLS AND METHODS

3.1 Experimental Setup

Our first experiment, *AATWordNet* was a replication of our mapping evaluation described in Tordai et al. [16]. As in our earlier evaluation the inter-rater agreement was low, one of our objectives for this experiment was to increase agreement by improving the description of matching categories. As summarized in Table 3.1, in this experiment we asked 5 raters to evaluate a sample of 74 lexical mappings between the Getty's Art and Architecture Thesaurus (AAT)[2] and Princeton WordNet version 2.0. The lexical mapping was based on string matching between preferred and/or alternative labels of concepts.

In the second experiment, *GTTinstance*, we aimed to rule out lexical mappings to WordNet as the cause of low inter-rater agreement due to WordNet's ambiguous word senses. We chose a different set of mappings created by a different matching technique for our second experiment. Raters had to evaluate 70 mappings, which were created using instance-based matching between the Dutch Royal Library's Gemeenschappelijke Trefwoordenthesaurus (GTT) and Brinkman Thesaurus, two subject heading thesauri. Instance based matching is based on instances, in this case books commonly annotated by concepts from both vocabularies.

In our last experiment, *GTTlexical*, we wanted to study lexical mappings between less ambiguous vocabularies than WordNet, and determine whether evaluation is easier when the two vocabularies are from the same domain. In this experiment we had 5 raters evaluate 75 lexical mappings between GTT and Brinkman.

In each experiment we provided raters with written guidelines on how to categorize mappings which included descriptions of the mapping categories and example mappings. We asked raters to evaluate mappings using our evaluation tool into 7 different categories. Additionally, we asked raters to "think aloud" by explaining their choice of categories and their application of the guidelines, which we transcribed. We then calculated the inter-rater agreement measurements for 7 categories and for 2 categories by aggregating the original categories. We then performed detailed analysis of the evaluations and of the raters' comments.
We describe the matching categories, guidelines and vocabularies in more detail in the next section.

Table 1: Overview of the three evaluation experiments

Experiment	Vocabularies	Mapping technique	# of raters	# of Mappings
AATWordNet	AAT and WordNet	lexical matching	5	74
GTTinstance	GTT and Brinkman	instance based matching	5	70
GTTlexical	GTT and Brinkman	lexical matching	5	75

3.2 Tools and Methods

3.2.1 SKOS relations and guidelines

We used the SKOS mapping properties to categorize the type of alignments. The `exactMatch` and `closeMatch` properties make

[2] `http://www.getty.edu/research/conducting_research/vocabularies/aat/`

statements about the degree of equality between two concepts. Hierarchical relations are expressed using `broadMatch` and `narrowMatch`, and `relatedMatch` expresses an associative relation between mapped concepts. In addition to these relations, we also defined a property to indicate that there is no relation between the mapped concepts: `unrelated`. We also give raters the option to choose "not sure" when they are unable to choose between the relations above.

As remarked earlier, we found in previous experiments [15, 16] that raters diverged greatly in the way they selected mapping properties. We attributed this divergence to an unclear differentiation between mapping relations. For example, we found in earlier experiments that raters varied greatly in their application of `relatedMatch`. One rater would find any relationship acceptable as related, while the other had a stricter definition. For this reason we wrote guidelines on the use of each mapping property for these experiments. Our rationale was to differentiate between each property as much as possible by describing them both in general terms, and by giving specific examples. For example, here is an excerpt from the guidelines on `relatedMatch`:

"***Related:****The two concepts have an associative relationship and are of two different (ontological) types. For example: a material and an object made from that material, such as milk and cheese, and an activity and object involved in the activity, such as the game volley ball and a volley ball. Generic examples of such relationships are: process and agent, action and property (e.g.; environmental cleanup and pollution), action and product (e.g.: weaving and cloth), cause and effect, object and origin, material and object, and object and practitioner.*"

We also defined the difference between `exactMatch` and `closeMatch`, and instructed raters to use `closeMatch` when the two concepts share the same label, but their parent concepts are different, as the vocabularies have different organizational schemes. An example of a `closeMatch` is *blowgun* where in one vocabulary it is a conduit and in the other it is a weapon. The two vocabularies present blowgun in different views: a structural view versus a functional view. The full guidelines can be found online[3].

In typical alignment evaluation settings researchers are interested in equality relations between concepts. In such cases the evaluation categories are equivalent and non-equivalent. Although we did not perform separate experiments with these categories we reduce the number of categories into two by summing up ratings of `exactMatch` and `closeMatch` into the equivalent category and the remainder into the non-equivalent category.

3.2.2 Vocabularies and mappings

In our *AATWordNet* experiment we use a sample of mappings between AAT and Wordnet. We generated the mappings[16] using string matching on preferred and alternative labels. AAT is an NISO standard compliant vocabulary [1] that we converted to SKOS, where each concept has preferred and alternative labels and is often accompanied by scope notes. WordNet contains synonymous labels grouped into synsets with no distinction between preferred and alternative labels. The meaning of each synset is clarified by glosses containing example sentences or definitions. Because in WordNet multiple synsets may share the same label, many of the lexical mappings between AAT and WordNet are ambiguous.

For our *GTTinstance* and *GTTlexical* experiments we used samples of mappings between the GTT and Brinkman thesaurus. Both thesauri are subject-heading vocabularies used to annotate the Dutch Royal Library's book collection and both include not only general descriptors but also geographic terms. Thus, the two vocabularies have the same purpose, although they differ in size and granularity: GTT is five times larger than Brinkman. Brinkman contains 13,025 concepts while GTT contains 65,297 concepts. The mappings we used in the second experiment were created as part of the STITCH[4] project using an instance-based matching technique described by Isaac et al. [10]. The sample of mappings we evaluated has no linguistic similarity, as all concepts with matching labels had been filtered out. For the third experiment we generated mappings between the two thesauri through lexical comparison of concept labels.

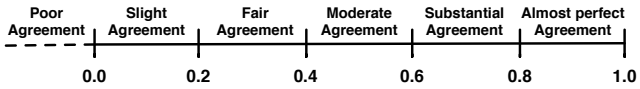

Figure 1: Interpretation of kappa values according to Landis and Koch. The scale is from -1 to 1

3.2.3 Interrater Agreement Measures

The simplest method for measuring agreement between raters is the percentage of agreement: *observed agreement*. Unfortunately it is difficult to interpret and compare across multiple experiments [4, 2], because it does not take into account agreement that occurs by chance. A number of measures exist that do correct for chance agreement. Cohen's kappa is used to measure agreement between two raters on nominal data. Fleiss' kappa is a generalization of Scott's pi and measures agreement between multiple raters on large sample sizes. Krippendorff's alpha, a more versatile measure, can be used with nominal, ordinal and interval type categories and even with missing data (for example when raters select "not sure"). Weighted Cohen's kappa allows us to count disagreements differently by using a weight matrix. The latter can be used, for instance, to count disagreement between `exactMatch` and `unrelated` heavier than a disagreement between `exactMatch` and `closeMatch`. All agreement measures use the observed agreement (or disagreement in the case of Krippendorff's alpha), that is the number of times raters agree, and an estimate of what the agreement would be if raters had assigned categories randomly.

These measures have two known problems: prevalence bias and annotator bias. Prevalence bias occurs when data falls into mostly one category: even if observed agreement between raters is high, the agreement measure may turn out low. In order to have a high measure the raters must agree on rare categories. Annotator bias occurs when the distribution of disagreement is highly skewed, leading to lower measures than when disagreements are more uniformly distributed.

While these measures are widely used, their interpretation is not clear cut. Landis and Koch [12] suggest the set of intervals displayed in Figure 1 based on their personal opinion. In linguistic research kappa values higher than 0.8 are considered acceptable [2]. The nature of the categorization task and the degree to which the matching categories can be defined operationally are factors in determining the minimum level of agreement for the results to be conclusive.

3.2.4 Evaluation Tool

We used an evaluation tool shown in Figure 2 to support raters in their judgment. Its user interface presents mappings with context information, such as the concept hierarchy, labels and scope notes. Raters can select mapping relations between concepts. The result-

[3] `http://www.cs.vu.nl/~atordai/Guidelines.pdf`

[4] `http://www.cs.vu.nl/STITCH/`

ing choices are stored in RDF using the OAEI alignment format [7] along with provenance information. This tool is a newer version of the one we used in Tordai et al. [15] . It is open source and available for download[5].

4. EXPERIMENTAL RESULTS

4.1 Quantitative Results

Table 2: Inter-rater agreement table for 7 mapping categories and for 2 categories. Measures include observed agreement between raters, the average of Cohen's kappa measured between each pair of raters, Fleiss' kappa over all raters and Krippendorff's alpha over all raters

Experiment	Observed agreement	Avg. Cohen's κ	Fleiss' κ	Krippendorff's α
	7 categories			
AATWordNet	0.69	0.564	0.565	0.575
GTTInstance	0.72	0.606	0.604	0.617
GTTlexical	0.85	0.473	0.475	0.475
	2 categories			
AATWordNet	0.84	0.666	0.669	0.679
GTTInstance	0.94	0.706	0.698	0.699
GTTlexical	0.95	0.514	0.538	0.538

The inter-rater agreement measures for our three experiments are displayed in Table 2. The AATWordNet experiment is a replication of the experiment described in [16] where we measured Cohen's kappa between pairs of raters and reported average measures of 0.36 for 7 categories and 0.67 for 2 categories with three raters. In the AATWordNet experiment we have an average Cohen's kappa of 0.564 for 7 categories which is a considerable increase over 0.36. We attribute this increase to a better description of the mapping categories in the guidelines, as we used the same evaluation tool and had five raters instead of three. There was however no improvement in the average Cohen's kappa for 2 categories, which suggests that while the guidelines were improved in the separate description of mapping categories, they did not help raters in making a distinction between equality and inequality more than in the previous evaluation.

Overall, we found higher agreement measures for 2 categories than for 7 categories which suggests that it is easier for raters to reach agreement over fewer categories.

Table 3: Distribution of SKOS matching relations used by raters in percentages. The ratings in each category are summed over all 5 raters

Experiment	exact	close	broad	narrow	related	unrelated	not sure
AATWordNet	31.9	6.5	4.1	5.4	10.8	40.8	0.5
GTTinstance	**5.7**	**3.7**	7.7	4.6	39.1	38.3	0.8
GTTlexical	**85.0**	9.1	1.0	1.8	1.0	2.1	0.0

The inter-rater agreement in the GTTlexical experiment is the lowest of all our experiments, despite the highest observed agreement (0.85). This is caused by prevalence bias, as 85% of the ratings fall into the exact-match category (see Table 3). The prevalence of one category causes the disagreement on the rare categories to weigh more heavily when measuring agreement. In the other experiments the distribution in the use of relations is less extreme than in GTTlexical.

We also found that the value of Fleiss' kappa is close to the average Cohen's kappa in all three experiments. Krippendorff's alpha is a bit higher for both AATWordNet and GTTInstance experiments because it does take into account missing values, in this case the use of the "not sure" category (see Table 3 for the distribution of mapping relations per experiment). In the GTTlexical experiment, where the "not sure" category was not used by raters, the value of Krippendorff's alpha is equal to the value of Fleiss' kappa.

Table 4: Cohen's kappa between each pair of raters for 7 categories from the GTTinstance experiment. The highest and lowest agreement is displayed in bold

Cohen's κ	Rater 1	Rater 2	Rater 3	Rater 4
Rater2	0.736	none		
Rater 3	0.534	0.665	none	
Rater 4	0.634	0.566	**0.483**	none
Rater 5	0.577	0.592	0.491	**0.783**

We measured Cohen's kappa for each pair of raters. Table 4 shows the values for the GTTinstance experiment with the highest value between Rater 4 and Rater 5 (0.783) and the lowest between Rater 3 and Rater 4 (0.483). The large difference is due to Rater 3's tendency to select related match for mappings that Rater 4 and 5 considered to be unrelated. We found a similarly high variation in the Cohen's kappa in AATWordNet and GTTlexical experiments which leads us to conclude that two raters are not enough to provide a consistent evaluation result.

Table 3 shows that in each experiment raters selected categories in very different distributions. In the GTTInstance experiment raters rarely selected the exact or close match categories. This was caused by the mapped concepts having no labels in common, as such mappings had been filtered out, therefore equivalent mappings were rare.

Table 5: The matrix of relations is the sum of the coincidence matrices from each experiment. The matrix shows the number of pairs of rating used by two raters for the same mapping. We consider it worse when raters mark opposing categories such as "exact" and "unrelated" than the categories "exact" and "close". The numbers are percentages of the total amount of observations in the three experiments (4420) and the numbers in bold represent agreements

Category	exact	close	broad	narrow	related	unrelated	not sure
exact	**37.30**						
close	4.88	**2.17**					
broad	0.59	0.86	**2.22**				
narrow	1.72	1.17	0.04	**1.63**			
related	1.54	1.36	0.68	0.68	**11.10**		
unrelated	0.63	0.72	1.49	0.81	5.84	**21.82**	
not sure	0.13	0.04	0.04	0.09	0.32	0.27	**0**

We examined the judgment of raters focusing on disagreements. The interrater-agreement measures when used on nominal data as-

[5] http://semanticweb.cs.vu.nl/amalgame/

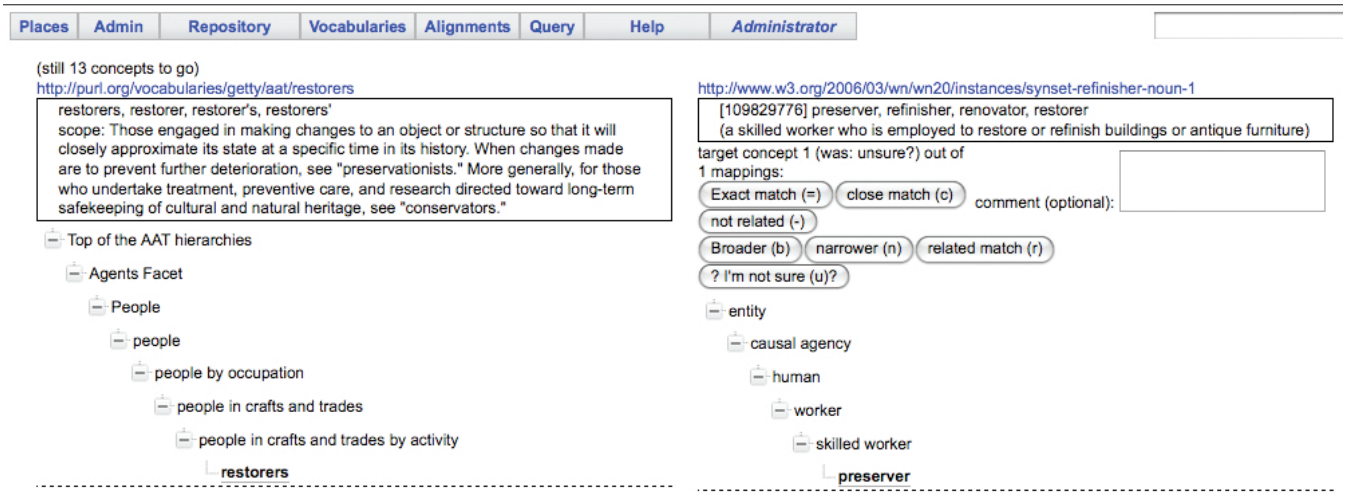

Figure 2: A partial screenshot of the tool evaluation used by raters. The screenshot shows a mapping between the concept *restorers* from AAT and *preserver* from WordNet. The labels and scope-notes are found in the upper boxes. This mapping caused high disagreement between raters

Table 6: Partial list of disagreements in mapping categories ordered by total number of occurrences in the three experiments. The total number of disagreements is 1068 and the number in parentheses is the percentage of total disagreements

	Categories disagreed upon	Occurences (%)
1.	related-unrelated	258 (24.39)
2.	exact-close	192 (20.42)
3.	exact-narrow	76 (7.18)
4.	exact-related	68 (6.43)
5.	broad-unrelated	66 (6.24)
...	...	...
10.	close-unrelated	32 (3.02)
..	...	...
12.	exact-unrelated	28 (2.65)
..	...	...
18.	broad-narrow	2 (0.19)

sume independence between categories. However, intuitively it is worse if raters disagree whether a mapping is an exact match or unrelated, than when they disagree on whether it is related or unrelated. Table 5 shows a matrix of coincidence of relations summed over all experiments. The agreements are along the diagonal, while the disagreements occupy the other cells. In Table 6 we isolated the pairs of relations raters disagreed upon, and ordered them according to the number of times they occurred in the experiments. The table shows that disagreements that can be considered the least "harmful" (i.e., the smallest semantic distance between the mapping relations involved) are the most frequent, such as the disagreements for related-unrelated and for exact-close. An analysis of the cases where raters selected "opposed" categories for the same mapping showed that it was mostly caused by one rater making a mistake.

Our main observation is that the inter-rater agreement measures are stable across our experiments. Although the inter-rater measures are relatively low, our analysis showed that most disagreements between raters are of the less "harmful" type.

4.2 Qualitative Results

We analyzed the use of SKOS matching relations by raters along with the reasons raters gave for their choices transcribed in the "think-aloud" sessions.

We found that overall raters selected different types of relations for lexical mappings than for non-lexical mappings. In the AAT-WordNet experiment mapped concepts share at least one label. Table 3 shows that most of the mappings were rated as either exact match or unrelated while the hierarchical relations (broader and narrower) were least frequently used. When mapped concepts are not equivalent they are either polysemes or homonyms. Homonyms have labels with the same spelling but the concepts unrelated (eg: *bends*; the act of bending and the decompression sickness). Polysemes are terms with different but related meanings, such as *milk*, being the product and the act of milking. In our experiments raters found polysemes difficult to rate because the boundary between relatedness and unrelatedness is not clear. In particular in WordNet, concepts are specifically divided into word senses thus distinguishing between various polysemic and homonymic forms. As a result, when raters evaluate mappings to WordNet they are confronted with multiple related and unrelated word senses. In our AATWordNet experiment raters found the concept *flow* from AAT, one of the most difficult to evaluate because it was mapped to 14 word senses in WordNet. This problem of distinguishing between meanings is not restricted to polysemes. In the GTTinstance experiment raters had the most difficulty in deciding whether mapped concepts were related or unrelated. For example, some raters found the concept *arid, dry territory* related to *erosion*, while others thought the link too remote to be useful. The fuzziness of concept boundaries and category boundaries makes agreement in evaluation more difficult to achieve. They are a manifestation of Lakoff's prototype theory where concepts far from the prototype become more difficult to categorize.

We also found that the contextual information such as hierarchy, multiple labels and scope note, can increase the difficulty in judgments in particular if they are contradictory. For example, the categorization of the mapping of concept *mantel* between AAT and WordNet, both referring to the thing around a fireplace, was complicated by the AAT scope note "Decorative frames around fire-

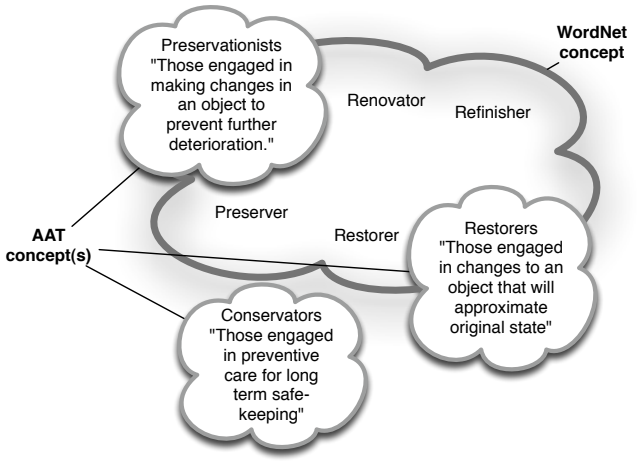

Figure 3: Fuzzy boundaries between AAT and WordNet concepts. The WordNet concept of *preserver* overlaps with multiple concepts from AAT through its labels: *Preservationists* and *Restorers*, but not with *Conservators*

place openings" and the WordNet gloss "shelf that projects from wall above fireplace". Three out of 5 raters judged the mapping "exact match". The fourth rater judged it "related" because the AAT parent concept is "furniture component" while the WordNet parent concept is "shelf" . The fifth rater judged the AAT term as broader as she considered the frame in the AAT scope note to be a broader than a shelf.

In comparison to AATWordNet, the GTTlexical experiment was judged "easier" by raters, as they were confronted with very few ambiguous mappings and made quicker judgments. In addition, both GTT and Brinkman contain few alternative labels and scope-notes limiting the amount of contextual information. In the GTTlexical experiment, where mapped concepts had the same label, raters tended to select exact match due to lack of context. In the GTTinstance experiment, however, the lack of context meant that concepts that were related were sometimes rated as unrelated by some raters. Both GTT and Brinkman vocabularies cover a wide range of subjects from "general culture" to economy, physics, history and even medicine. Thus evaluating concepts from more specific domains was more difficult due to lack of context. For example, the mapping between the drug *Dapsone* and the disease *leprosy* was rejected by some raters because the parent concept of Dapsone is *anti-epileptic drug*. Other raters looked up Dapsone on Wikipedia, found that it is also a drug used for leprosy and selected related match because the therapy and disease have an associative relationship. In this case, if we had prohibited the use of outside sources such as Wikipedia, all raters would have most likely selected unrelated based on the available information (unless one of them was a medical expert) which would have led to higher inter-rater agreement. However, a related match between the two concepts can be useful in some applications. Our experiments have shown that raters behave differently and some are more inclined to look up information than others.

We found that for some mappings raters thought the SKOS matching relations inadequate. Although raters could use the unrelated category whenever the mapping was not a SKOS relation, they were reluctant to reject mappings with some semantic link. In particular in the AATWordNet and GTTlexical experiments some aligned concepts (partially) overlapped each other in meaning, therefore warranting some sort of equivalence relation that could not be defined as exact or close match. An example from AATWordNet is the concept *restorer* as shown in Figure 3. The WordNet concept for restorer also included the labels "refinisher", "renovator" and "preserver", whereas *restorers* and *preservationists* were separate concepts in AAT. Raters were reluctant to reject the mapping from AAT's *restorer* to WordNet's *restorer*, but felt neither exact nor close match was appropriate.

A complementing explanation of possible differences between raters could be based on the variability of subjective guidelines that raters appear to construct during the evaluation task. This view is supported by the notion of situated cognition [5] that stipulates that people construct their knowledge "on the fly" in a specific context. When raters were confronted with a non-prototypical mapping they formed their own interpretation of the guidelines and applied that particular rule to similar mappings. For example, one rater created the following rule during the GTTinstance experiment: "*if two concepts are not on the same level of specialization they cannot be related*". The rater continued to apply this rule throughout the evaluation, even though our guidelines did not contain such a specific rule, and no other raters formulated it so clearly.

The background of raters also had an impact on their process of categorization. Two of the raters had a strong thesaurus background that influenced some of their choices. For example, one of these raters would not categorize mappings as broad or narrow match if they shared the same label, commenting that it is not proper ISO standard practice [1]. Raters without this background had no reservations in using hierarchical categories on mappings that shared the same label. We did not specify a purpose or task for the alignments but it seems that raters with a thesaurus background thought of the mappings in terms of a thesaurus merging task, while other raters thought of mappings in terms of an annotation task. Our findings are similar to those reported by Bailey et al. [3] in the field of information retrieval, where they found that judges with different levels of specialization in the task had low agreement.

In the experiments we found that certain disagreements are caused by different interpretations of differences in thesauri, in particular in their hierarchy. GTT is organized according to is-a type relations, while in Brinkman concepts are organized according to a mix of is-a and part-of relations. This sometimes caused problems when the relation a rater wanted to choose contradicted the relation in the thesaurus. For example, there were disagreements on the mapping between *Waste products* from Brinkman to *Environmental pollution* in GTT. Some raters chose related match in accordance with our guidelines about cause and effect, while other raters chose broader because in Brinkman *Environmental pollution* is the parent concept of *Waste products*. Such disagreements can be avoided by adding additional guidelines, but in practice it is often impossible to foresee the effect specific differences between thesauri have on evaluation.

5. DISCUSSION AND CONCLUSION

Manual evaluation is a method for establishing the quality of vocabulary alignments. High agreement between raters is a requirement for being able the make conclusive statements about quality of alignment methods. However, there are a number of factors that influence the judgment of raters. In this paper we studied the process of manual evaluation and found that there are aspects of the evaluation setup that can be controlled, and aspects that make the task inherently difficult.

One aspect that can be controlled is the provision of clear guidelines to the evaluators. Guidelines should include clear examples, precise descriptions of the categories, and instructions how to deal

with thesaurus errors. The granularity of the categories is another factor where choices can be made. In our experiments we chose to use the SKOS mapping categories, but our results show that a two category system (match/no match) leads to higher reliability measures. On the other hand, some of our raters indicated that they found the SKOS categories too limited.

Another aspect that can be controlled is the nature of the sample. One can simply choose to select a random sample of alignments or one can construct a sample that contains certain types of alignments as we did in the GTTinstance evaluation. Although our results indicate that different samples lead to similar values for inter-rater reliability, the choice of a specific sample can circumvent certain problems such as prevalence bias.

Aspects largely beyond control are lexical ambiguity and rater characteristics. It is well known from studies in lexical semantics that the boundary between polysemy and homonomy is vague and that the classification of different types of polysemy is still a matter of debate among linguists. Humans rarely have problems with disambiguating the meaning of words in a discourse context. However, in an ontology alignment task this context is usually much more limited than in discourse.

The evaluation process can also be influenced by the background of the evaluators of alignments. Domain specialists (e.g. in a medical or cultural heritage domain) may use different evaluation criteria than raters with a linguistic or a computer science background. Of course one can choose to select raters with a similar background.

A related factor is the purpose of the ontology alignment. For example, if the aligned concepts are used to retrieve documents annotated by different ontologies in the domain of medicine, the difference between organs of a human and a mouse may not be of great importance. In other applications such differences may be essential. Of course the guidelines can be adapted to the nature of the application, but this make comparison of the quality of alignment methods much more complex.

In summary, our results indicate that the manual evaluation of ontology alignments is by no means an easy task and that the ontology alignment community should be careful in the construction and use of reference alignments. We recommend that the OAEI community starts establishing best practices and guidelines for constructing reference alignments. Based on this paper we suggest to include at least the following elements in such an evaluation methodology:

- Select one of the three interrater-agreement measures used in Table 2 as the prescribed standard. Although from the results reported in this paper there is no clear winner, we suggest using Krippendorff's alpha for its versatility as it can be used with any number of raters, with incomplete data and on different sample sizes. The use of an inter-rater agreement measure will make comparison between experiments of different authors easier.

- Prescribe a minimum set of raters for manual evaluation. This minimum should not lower than 3. A range of 3-5 raters appears reasonable.

- Agree on a set of alignment relations. The SKOS relations are attractive candidates mainly because they are part of an heavily-used standard for publishing thesauri on the Web. However, the set of equivalence relations used by Halpin et al. [9] has a more formal underpinning.

- Agree on a set of guidelines for helping to decide which mapping relation to use. The guidelines provided by us (cf. `http://www.cs.vu.nl/~atordai/Guidelines.pdf`) might serve as a place to start.

Having said this, we agree with Lakoff's view on categorization and its consequence: we should not expect full agreement on reference alignments. On the other hand, we expect many Web applications to be able to live with this and still make useful semantic links.

Acknowledgements

We would like to thank our raters: Mark van Assem, Victor de Boer, Marieke van Erp, Michiel Hildebrand, Veronique Malaisé, Lourens van der Meij, Carmen Reverté and Roxane Segers for their participation in our experiments. This research was supported by the MultimediaN project funded through the BSIK programme of the Dutch Government.

6. REFERENCES

[1] Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Technical report, National Information Standards Organization, 2005.

[2] R. Artstein and M. Poesio. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34:555–596, 2008.

[3] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance Assessment: Are Judges Exchangeable and Does It Matter?. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 667–674. ACM, 2008.

[4] J. Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22:249–254, June 1996.

[5] W. J. Clancey. *Situated Cognition: On Human Knowledge and Computer Representations*. Cambridge University Press, New York, NY, USA, 1997.

[6] J. Cohen. A Coefficient of Agreement For Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960.

[7] J. Euzenat. An Api for Ontology Alignment. In S. A. McIlraith, D. Plexousakis, and F. van Harmelen, editors, *The Semantic Web – ISWC 2004*, volume 3298 of *Lecture Notes in Computer Science*, pages 698–712. Springer Berlin / Heidelberg, 2004.

[8] J. Euzenat, A. Ferrara, L. Hollink, A. Isaac, C. Joslyn, V. Malaisé, C. Meilicke, A. Nikolov, J. Pane, M. Sabou, F. Scharffe, P. Shvaiko, V. Spiliopoulos, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, C. Trojahn dos Santos, G. Vouros, and S. Wang. Results of the Ontology Alignment Evaluation Initiative 2009. In P. S. et al., editor, *Proc. 4th ISWC workshop on ontology matching (OM)*, pages 73–126, 2009.

[9] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuiness, and H. S. Thompson. When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data. In *Proceedings of the 9th International Semantic Web Conference (ISWC)*, November 2010.

[10] A. Isaac, L. Van Der Meij, S. Schlobach, and S. Wang. An Empirical Study of Instance-based Ontology Matching. In *ISWC'07/ASWC'07*, pages 253–266, Berlin, Heidelberg, 2007. Springer-Verlag.

[11] G. Lakoff. *Women, Fire and Dangerous Things; What Categories Reveal About the Mind*

[12] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174, 1977.

[13] A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System Reference, August 2009.

[14] J. R. Taylor. *Linguistic Categorization*. Oxford University Press, third edition, 2003.

[15] A. Tordai, J. van Ossenbruggen, and G. Schreiber. Combining Vocabulary Alignment Techniques. In Y. Gil and N. F. Noy, editors, *K-CAP*, pages 25–32. ACM, 2009.

[16] A. Tordai, J. van Ossenbruggen, G. Schreiber, and B. J. Wielinga. Aligning Large SKOS-like Vocabularies: Two Case Studies. In *The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3, 2010, Proceedings, Part I*, volume 6088 of *Lecture Notes in Computer Science*, pages 198–212. Springer, 2010.