Two Variations on Ontology Alignment Evaluation: Methodological Issues

Laura Hollink, Mark van Assem, Shenghui Wang, Antoine Isaac, Guus Schreiber

Department of Computer Science
Vrije Universiteit Amsterdam
de Boelelaan 1081 HV
The Netherlands

Abstract. Evaluation of ontology alignments is in practice done in two ways: (1) assessing individual correspondences and (2) comparing the alignment to a reference alignment. However, this type of evaluation does not guarantee that an application which uses the alignment will perform well. In this paper, we contribute to the current ontology alignment evaluation practices by proposing two alternative evaluation methods that take into account some characteristics of a usage scenario without doing a full-fledged end-to-end evaluation. We compare different evaluation approaches in three case studies, focussing on methodological issues. Each case study considers an alignment between a different pair of ontologies, ranging from rich and well-structured to small and poorly structured. This enables us to conclude on the use of different evaluation approaches in different settings.

1 Introduction

The rise of the semantic web has lead to a large number of different and heterogeneous ontologies. This has created a need to interconnect these ontologies. Tools and algorithms have emerged that automate the task of matching two ontologies (see e.g. [18] or [13] for an overview). They create sets of correspondences between entities from different ontologies, which together are called alignments [6]. These correspondences can be used for various tasks, such as ontology merging, query answering, data translation, or for navigation on the semantic web[1]. Although the performance of ontology matching tools has improved in recent years [9], the quality of an alignment varies considerably depending on the tool and the features of the ontologies at hand.

The need for evaluation of ontology alignments has been recognised. Since 2004, the Ontology Alignment Evaluation Initiative (OAEI) organizes evaluation campaigns aimed at evaluating ontology matching technologies[2]. This has lead to the development of mature alignment evaluation methods. In this paper, we

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

contribute to evaluation practices by investigating two alternative evaluation strategies.

Evaluation of alignments is commonly done in two ways: (1) assessing the alignment itself by judging the correctness of each correspondence and (2) comparing the alignment to a gold standard in the form of a reference alignment. However, this type of evaluation does not guarantee that an application which uses the alignment will perform well. Evaluating the application that uses the alignment – commonly referred to as end-to-end evaluation – will provide a better indication of the value of the alignment [20]. In one of the seven test cases in the 2007 OAEI an end-to-end evaluation was performed; alignments were evaluated on the basis of their value for an annotation translation scenario [11].

Although many agree that end-to-end evaluation is desirable, it is hard to realize in practice. Even in a large scale initiative like OAEI it is time consuming. Another complicating factor is that real-world applications that use alignments are as yet scarce, and associated data on user behaviour and user satisfaction is even more rare. A more feasible alternative is to take into account some characteristics of a particular usage scenario without doing a full-fledged end-to-end evaluation [12]. The OAEI more and more incorporates usage scenarios in the evaluation. For example, in the Anatomy track of OAEI 2007, tools were asked to return a high-precision and a high-recall alignment, supporting the respective usage scenarios of fully automatic alignment creation and suggestion of candidate alignments to an expert [9]. Also, the number of tracks and test cases has increased every year [7, 8, 9], recognising the need for matching ontologies with different features, such as size, richness and types of relations (e.g. rdfs:subClassOf, part-of), depth of the hierarchy, etc.

Continuing this line of research, we investigate two evaluation methods that approximate end-to-end evaluation. Each strategy takes into account one characteristic of the targeted application context; the first takes into account the expected frequency of use of each correspondence and the second considers the expected effect of a particular misalignment on the performance of the application. We compare these two alternatives to the more common evaluation methods of assessing each individual correspondence and comparison to a reference alignment. We perform three case studies in which we evaluate an alignment using the four methods. This enables us to discuss not only the practicalities of each approach, but also the different conclusions that are the outcome of the approaches. Each case study compares a different pair of ontologies. Since the ontologies are structured differently, this allows us to discuss the effect of the features of the ontologies on the types of alignment errors that occur. In addition, we perform an end-to-end evaluation and compare this to the outcome of the four evaluation methods in a qualitative way.

The purpose of the paper is twofold. First, we intend to gain insight in methodological issues of evaluation methods. Second, we intend to give insight into the effect of the characteristics of the ontologies on the quality of the alignment, and on the best evaluation method to choose. It is not the purpose of this paper to evaluate a particular alignment or matching tool. In Section 2

we discuss the proposed methods, together with related work concerning these methods. In Section 3 we present our case studies and in Section 4 the end-to-end evaluations. In Section 5 we discuss the results.

2 Alignment Evaluation Methods

Ideas have been put forward to find feasible alternatives to end-to-end evaluation. In this section we discuss related work on this topic, and describe two methods that each take into account one characteristic of an application that uses an alignment. The application scenario that we focus on is a query reformulation scenario, in which users pose a query in terms of one ontology in order to retrieve items that are annotated with concepts from another ontology. We assume that there is a partial alignment between the two ontologies, which is a realistic assumption given the state-of-the art of matching tools.

2.1 Evaluating most frequently used correspondences

If an alignment is large, evaluating all correspondences can be a time consuming process. A more cost-effective option is to evaluate a random sample of all correspondences, and generalize the results to get an estimate of the quality of the alignment as a whole.

An alternative to taking a random sample is purposefully selecting a sample. Van Hage et al. [20] note that in a particular application some correspondences affect the result (and thus user satisfaction) more than others. An end-to-end evaluation can take this into account but an evaluation of all (or a sample of) individual correspondences cannot. Evaluating the most important correspondences would better approximate the outcome of an end-to-end evaluation. The notion of 'importance' can mean different things in different application contexts. In this paper, we propose to use the estimated frequency of use of each correspondence as a weighting factor in the computation of performance measures. To this end, we divide all correspondences into two strata: infrequently and frequently used correspondences. As shown by Van Hage et al. [20], an intuitive way of stratified sampling is to aggregate the results of the strata 1 to L in the following manner:

$$\hat{P} = \sum_{h=1}^{L} \frac{N_h}{N} \hat{P}_h \tag{1}$$

where $\hat{P}$ is the estimated performance of the entire population, $\hat{P}_h$ is the estimated performance of stratum h, $\frac{N_h}{N}$ is a weighting factor based on the relative size of the stratum, where N_h is the size of the stratum, and N is the total population size.

Instead of weighting the strata based on their size, we propose to weight them based on their expected frequency of use:

$$\hat{P} = \sum_{h=1}^{L} \frac{\sum_{a \epsilon H} \text{freq}(a)}{\sum_{a \epsilon A} \text{freq}(a)} \hat{P}_h \quad (2)$$

where freq(a) is the frequency of use of correspondence a, H is the total set of correspondences in stratum h, and A is the total set of correspondences in the alignment.

Selecting the most frequently used correspondences for evaluation is beneficial in two situations. First, if there is a difference in quality between the frequently used correspondences and the infrequently used correspondences, the frequency-weighted precision will give a more reliable estimate of the performance of the application using the alignment. Second, if one intends a semi-automatic matching process in which suggested correspondences are manually checked and corrected by an expert, the frequency provides an ordering in which to check. This kind of scenario is targeted in the Anatomy track of OAEI 2007 [9] by asking participants to generate a high-recall alignment. Ehrig and Euzenat [4] consider the semi-automatic matching process by measuring the quality of an alignment by the effort it will take an expert to correct it. We argue that correction of a number of frequently used correspondences will positively affect the performance of the application more than correction of the same number of randomly selected correspondences.

Implementation of the method To estimate the frequency of use of each correspondence, we assume that each concept in source ontology X has an equal probability of being selected as a query by a user. For each query concept x, we determine the closest concept x' in X that has a correspondence to a concept y in ontology Y (the target ontology with which items are annotated). Closeness is determined by counting the number of steps in the (broader/narrower) hierarchy between x and x'. If a query concept x does not itself have a correspondence to Y, the correspondence of x' to ontology Y is used to answer the query, thus adding to the frequency count of correspondence $\{x', y\}$. Our estimation is biased, because in practice some query concepts are more often used than others. Logs of user and system behaviour can be used to determine more accurate prior probabilities of each query concept. However, logs are not always available.

2.2 Semantic distance to a reference alignment

Comparing an alignment A to a reference alignment R gives precision as well as recall scores. Precision is the proportion of correspondences in A that are also found in reference alignment R, while recall is the proportion of the reference alignment R that is covered by A.

Incorrect correspondences negatively affect the performance of an application. However, this effect varies depending on how incorrect the correspondence is. Performance of an application will drop steeply if a correspondence links two completely unrelated concepts, while it may drop only slightly if a correspondence links two closely related concepts. We investigate the use of a semantic

distance measure to capture this difference. More specifically, we use semantic distance to represent the distance between a correspondence in A and a correspondence in a reference alignment R. This allows us to distinguish between correspondences that cause incorrect results, and correspondences that are misaligned but still produce an acceptable result in the application.

The idea of a more nuanced precision and recall measure has been proposed before. Ehrig and Euzenat [4] propose to include a proximity measure in the evaluation of alignments. They suggest to use the effort needed by an expert to correct mistakes in an alignment as a measure of the quality of an alignment. In the same paper, they propose to use the proximity between two concepts as a quality measure. A very simple distance measure is used as an example.

In the current paper, we implement this idea by using the semantic distance measure of Leacock and Chodorow [15]. This measure scored well in a comparative study of five semantic distance measures by Budanitsky and Hirst [3], and has the pragmatic advantage that it does not need an external corpus. The measure by Leacock and Chodorow, sim_{LC}, actually measures semantic proximity:

$$sim_{LC} = -\log \frac{len_{(c_1,c_2)}}{2D}$$

where $len_{(c1,c1)}$ is the shortest path between concepts $c1$ and $c2$, which is defined as the number of nodes encountered when following the (broader/narrower) hierarchy from $c1$ to $c2$. D is the maximum depth of the hierarchy.

In our case studies, we compare each correspondence $\{x, y\}$ in A to a correspondence $\{x, y'\}$ in a reference alignment R. We use the semantic distance between y and y' as a relevance measure for the correspondence $\{x, y\}$. To calculate precision and recall, we normalize the semantic distance to a scale from 0 to 1.

A side effect of using a semantic distance measure is that the assessments are no longer dichotomous but are measured on an interval level. Common recall and precision measures are not suited for this scale. Therefore, we use *Generalised Precision* and *Generalized Recall* as proposed by Kekäläinen and Järvelin [14]:

$$gP = \sum_{a \epsilon A} \frac{r(a)}{|A|} \qquad gR = \frac{\sum_{a \epsilon A} r(a)}{\sum_{a \epsilon R} r(a)} \tag{3}$$

where $r(a)$ is the relevance of correspondence a, A is the set of all correspondences found by the matching tool, and R is the set of all correspondences in the reference alignment. A similar notion of this measure was later described by Euzenat [5]. The latter measure is more general since it is based on an overlap function between two alignments instead of distances between individual correspondences.

3 Case Studies

In this section we employ the four evaluation methods to evaluate alignments between three pairs of vocabularies. In three case studies, one for each alignment, we discuss the different outcomes of the evaluation methods.

In each case study we take a different source vocabulary: SVCN, WordNet and ARIA. The target vocabulary is the same in each case study: the Art and Architecture Thesaurus (AAT). Strictly speaking, some of these vocabularies are not ontologies. In practice, however, many vocabularies can be seen as ontologies with less formal semantics. They are widely used as annotation and search vocabularies in, for example, the cultural heritage field. Ontology alignment tools can be applied to these vocabularies, although the looser semantics may influence the quality of the alignment. We used the RDF representations of AAT, SVCN and ARIA provided by the E-Culture project [17], and WordNet's RDF representation by the W3C [19]. The three source vocabularies differ in size, granularity, structure, and topical overlap with the AAT, which allows us to investigate the role of vocabulary features in the evaluation methods.

The Getty Institute's AAT[3] is used by museums around the world for indexing works of art [16]. Its concepts have English labels and are arranged in seven facets including Styles and Periods, Agents and Activities. In our study we concentrate only on the Objects facet, which contains 16,436 concepts ranging from types of chairs to buildings and measuring devices, arranged in a monohierarchy with a maximum depth of 17. The broader/narrower hierarchy of this facet is ontologically clean. Concepts in the AAT typically have many labels, e.g. "armchair", "armchairs", "chairs, arm", "chaises á bras". The RDF version provided by the E-Culture project also has Dutch labels, obtained from a translation of AAT.[4]

The alignments were made with Falcon-AO, an automatic ontology matching tool [10]. We selected Falcon-AO because it was among the best performing tools in the OAEI 2006. We employed it as an off-the-shelf tool, because this paper does not focus on *tool* evaluation but on the characteristics of different methods of *alignment* evaluation.

3.1 Case 1: Alignment between AAT and SVCN

SVCN is a thesaurus developed and used by several Dutch ethnographic museums.[5] It has four facets, of which the Object facet has 4,200 concepts (making it four times smaller than the Object facet in the AAT). SVCN's Object facet was originally created by selecting AAT concepts and translating the labels to Dutch. However, over time intermediate and leaf concepts have been inserted and removed, resulting in a hierarchy with a maximum depth of 13. The broader/narrower hierarchy is well-designed, but contains more errors than AAT's.

Evaluating individual correspondences Falcon produced 2,748 correspondences between SVCN and AAT. We estimated the frequency of use of each

[3] See `http://www.getty.edu/research/conducting_research/vocabularies/aat/`. The AAT is a licensed resource.

[4] `http://www.aat-ned.nl/`

[5] `http://www.svcn.nl/`

correspondence, as described in Section 2.1. Figure 1 displays cumulative percentages of these frequencies against cumulative percentages of the number of correspondences; all correspondences were ordered according to their frequency and displayed so that infrequent correspondences appear on the left side of the figure and the most frequent correspondences appear on the right. If each correspondence was used equally frequently, the graph would show a straight line from the origin to the top-right corner. SVCN does not deviate much from this straight line.

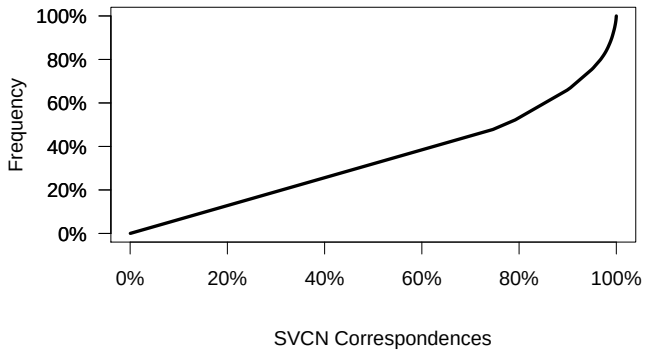

Fig. 1. Cumulative percentage of estimated use of SVCN-AAT correspondences in the application scenario. The total number of correspondences is 2,748.

All correspondences were divided over two strata: frequently used and infrequently used correspondences. The size of the frequent stratum was set to 80 (3% of all correspondences), which are responsible for 20% of the use in the application scenario (Figure 1). The choice for a size of 80 is pragmatic: it is a low number of correspondences that can be evaluated but still reflects a large frequency percentage. We evaluated all correspondences in the frequent stratum and a random sample of 200 from the infrequent stratum. Table 1 shows that the precision of the two strata differs, but not significantly so (0.93 and 0.89). We then weighted the outcomes of these evaluations in two ways: (1) according to the sizes of the strata (80 and 2,668) as in Equation 1 and (2) according to the frequency of use of the correspondences in the strata as in Equation 2. Both weighting schemes gave a precision of 0.89 (see Table 1).

Since in this use case the size of the population of all correspondences is large compared to the sample sizes, we used the binomial distribution to approximate the margins of error (shown in Table 1). The margin of error of a binomial distribution is given by:

$$\text{Margin of error} = 1.96\sqrt{\frac{p(1-p)}{n}} \tag{4}$$

Evaluation Type	Precision	Recall
Random sample of infrequent stratum	0.89±0.04	
Frequent stratum	0.93	
Weighted based on stratum size	0.89±0.03	
Weighted based on frequency of use	0.89±0.03	
After correction of frequent stratum	0.91±0.03	
After random correction	0.93±0.03	
Comparison to a reference alignment	0.84±0.07	0.80±0.08
Semantic distance to a reference alignment	0.90±0.06	0.86±0.07

Table 1. Evaluation of the alignment between SVCN and AAT.

One reason for taking into account the frequency of use of correspondences is that it gives an order in which to manually check and correct the correspondences. We corrected all 80 correspondences in the frequent stratum and then recalculated the precision of the alignment, weighted by frequency of use. This gave a precision of 0.91, which is not a significant increase. After manual correction of a random sample of 80 correspondences the precision rises to 0.93, which is higher but again not a significant increase. A possible reason for the finding that random correction gives a better precision than correction of frequent correspondences, is the fact that there were more wrong correspondences in the random sample. Another factor is that the contribution to the total frequency of correspondences in the two strata is similar.

Comparison to a reference alignment Reference alignment evaluation has the advantage that both precision and recall can be determined, but it is more costly because two vocabularies have to be aligned completely. Instead of aligning all concepts, we took a random sample of 100 concepts from SVCN and aligned those to AAT. Based on this partial reference alignment, Falcon's alignment has a precision of 0.84 and a recall of 0.80 (Table 1). As an alternative, we employ a semantic distance measure to compare the correspondences to the reference alignment; each correspondence $\{x,y'\}$ in the reference alignment is compared to a correspondence $\{x, y\}$ delivered by Falcon. We use the sim_{LC} measure between y and y', which results in a scaled value (0-1). Generalized precision and recall can then be calculated over these values (see Table 1). In the case of SVCN, the semantic distance based precision and recall are higher that the 'traditional' precision and recall, but the differences lie within the margins of error.

3.2 Case 2: Alignment between AAT and WordNet

WordNet is a freely available thesaurus of the English language developed by Princeton[6]. It has three top concepts: Physical entity, Abstraction and Thing. We only used the hierarchy below Physical entity ◁– Physical object, which contains 31,547 concepts. Each concept has multiple synonymous terms. The main hierarchy is formed by the polyhierarchic hyponym relation which contains more

[6] `http://wordnet.princeton.edu/`

ontological errors than AAT's hierarchy. The topical overlap with AAT is reasonable, depending on the part of the hierarchy. WordNet's Physical object hierarchy covers, for example, also biological concepts such as people, animals and plants while the Object facet of AAT does not. Other parts of WordNet are very similar to AAT. For example, the hierarchy from Furniture down to Chesterfield sofas is almost identical to that in AAT. The maximum depth of the Physical entity hierarchy is the same as AAT' Object Facet: 17 nodes.

Evaluating individual correspondences Falcon produced 4.101 correspondences between WordNet and AAT. Applying our frequency estimation gives a distribution depicted in Figure 2. In this case the contribution of the most frequent correspondences is much greater; the top 20% of correspondences is already responsible for 70% of expected usage (reminiscent of Zipf's law).

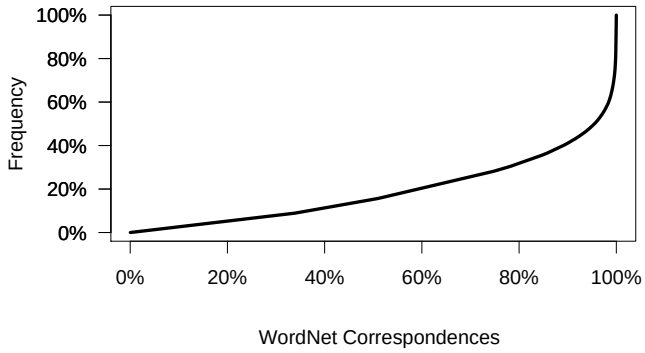

Fig. 2. Cumulative percentage of estimated use of WN-AAT correspondences in the application scenario. The total number of correspondences is 4.101.

We performed the same evaluation procedures as for the SVCN case, except that the size of the frequent stratum was set to 30 (0.7% of all correspondences). This is possible because here the contribution of the top correspondences is greater; the top 30 is responsible for 33% of total frequency. This reduction saves us a considerable evaluation effort. The results of the different evaluation methods are presented in Table 2. In the case of WordNet, weighting based on stratum size gives a slightly higher precision than weighting based on frequency (0.71 and 0.68, respectively).

Manual correction of all 30 frequent correspondences gives a higher precision than correcting 30 randomly selected correspondences from the complete set of correspondences (0.81 and 0.72, respectively, calculated by frequency-based weighting). This shows that in the WordNet case, it is sensible to prioritize correction of the most frequent correspondences.

Evaluation Type	Precision	Recall
Random sample of infrequent stratum	0.72±0.06	
Frequent stratum	0.60	
Weighted based on stratum size	0.71±0.05	
Weighted based on frequency of use	0.68±0.04	
After correction of frequent stratum	0.81±0.04	
After random correction	0.72±0.04	
Comparison to a reference alignment	0.62±0.10	0.45±0.10
Semantic distance to a reference alignment	0.64±0.09	0.47±0.10

Table 2. Evaluation of the alignment between WordNet and AAT.

Comparison to a reference alignment We performed a sample reference alignment evaluation in the same manner as for the SVCN case (n=100). The results are much lower than those for SVCN, as can be seen in Table 2. The margins of error are somewhat higher because the sample size is smaller. The effect of applying semantic distance is smaller than the effect we saw for SVCN.

3.3 Case 3: Alignment between AAT and ARIA

ARIA is a thesaurus developed by the Dutch Rijksmuseum for a website that showcases some 750 masterpieces of the collection.[7] It contains 491 concepts which are all art-related object types. There are 26 top concepts such as Altarpieces, Household scenes and Clothing, only half of which have subconcepts. Each concept has one term in Dutch and one in English. Its hierarchy is at most 3 concepts deep and is arranged in a polyhierarchy; e.g. Retables is subordinate to Altarpieces and Religious paraphernalia. The broader/narrower relation used in ARIA can in many cases not be interpreted as `rdfs:subClassOf`. For example, Costumes and textiles has a grandchild Portable altars. ARIA is the smallest and most weakly structured of the three source vocabularies.

Evaluating individual correspondences Falcon produced 278 correspondences between ARIA and AAT. Figure 3 shows the results of applying our frequency estimation. In this case the contribution of the most frequent correspondences is large; the top 20% of correspondences is responsible for 50% of expected usage.

Again we opted for a size of 30 for the frequent stratum (6% of all correspondences), which are responsible for 42% of the use in the application scenario. In this case, weighting according to the size of the stratum gave a precision of 0.74, while weighting according to the frequency of use gave a precision of 0.70.

Since the sample size is large compared to the size of the population of all correspondences in this case, we cannot approximate the margin of error with a binomial distribution. Instead, we used the following equation to compute the margin of error for a hypergeometric distribution [2]:

$$\text{Margin of error} = 1.96\frac{N-n}{N-1}m(1-\frac{m}{n}) \tag{5}$$

[7] `http://www.rijksmuseum.nl/aria/`

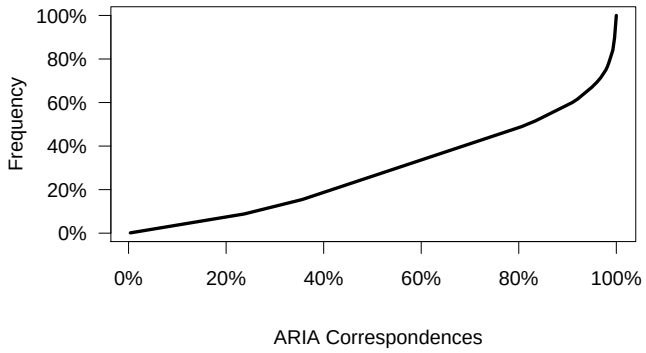

Fig. 3. Cumulative percentage of estimated use of ARIA-AAT correspondences in the application scenario. The total number of correspondences is 278.

where N is the size of the population, n is the size of the sample, and m is the number of correct correspondence found in the sample.

Evaluation Type	**Precision**	**Recall**
Random sample of infrequent stratum	0.75±0.03	
Frequent stratum	0.63	
Weighted based on stratum size	0.74±0.03	
Weighted based on frequency of use	0.70±0.03	
After correction of frequent stratum	0.85±0.02	
After random correction	0.74±0.03	
Comparison to a reference alignment	0.66±0.09	0.63±0.09
Semantic distance to a reference alignment	0.80±0.08	0.76±0.08

Table 3. Evaluation of the alignment between Aria and AAT.

Manual correction of the most frequent stratum gives a precision of 0.85, which is again higher than random correction (0.74).

Comparison to a reference alignment We performed a sample reference alignment evaluation in the same manner as for SVCN and WordNet (n=100). The recall and precision measures as shown in Table 3 are in between those for SVCN (highest) and WordNet (lowest). The effect of applying semantic distance is considerable.

4 End-to-end evaluation

In this section we present an end-to-end evaluation performed using the three alignments from the case studies. The application scenario that we focus on is

a query reformulation task for information retrieval: a user query for concept $x \in$ vocabulary X is transformed into a concept $y \in$ vocabulary Y. We queried a dataset of 15,723 art objects indexed with AAT provided by the E-Culture project. Objects annotated with concepts from Y are returned to the user and the relevance of these objects to the query x is rated. We used 20 randomly selected query concepts from each source vocabulary [8] and evaluated two different methods of reformulation for cases where a query $x \in X$ has no direct correspondence to Y: (1) find a concept x' in the hierarchy below x that has a correspondence to a concept $y \in Y$ (strategy "below"); or (2) find a concept x' above x with a correspondence to a concept $y \in Y$ (strategy "above").

The effectiveness of the reformulation was evaluated by assessing the relevance of objects annotated with concept y (or subconcepts of y) on on a six-point scale ranging from "very relevant" to "not relevant at all". Generalized precision and recall were calculated from these ordinal assessments (see Table 4). For comparison we also calculated precision and recall based on dichotomous (0/1) assessments by rescaling the ordinal values 0-2 to 0 and 3-5 to 1. Recall was calculated based on a recall pool[9].

		Precision		**Recall**	
Vocabulary	**Strategy**	**Binary Scale**	**6-point Scale**	**Binary Scale**	**6-point Scale**
ARIA	upward	0.27	0.37	0.83	0.88
	downward	0.70	0.66	0.49	0.43
SVCN	upward	0.46	0.48	0.93	0.96
	downward	0.79	0.76	0.42	0.36
WordNet	upward	0.46	0.48	0.80	0.81
	downward	0.63	0.67	0.18	0.18

Table 4. Precision and Recall of end-to-end evaluation for a six-point scale and a binary scale. Results are shown for two different query reformulation strategies.

We stress that the precision and recall figures presented in Table 4 refer to *relevancy* of the returned objects, instead of *correctness* of correspondences. This means that it is not possible to directly compare the results from reference alignment evaluations with the results from end-to-end evaluation. This is a general methodological difficulty when comparing evaluation methods, not only for the scenario presented in this paper.

5 Interpretation and Discussion of Results

The three case studies have illustrated differences between the evaluation methods and between the aligned vocabularies. The results show that the different evaluation methods stress different properties of an alignment.

[8] We excluded concepts that were too general such as Physical object

[9] A recall pool consists of the union of all objects returned by any of the systems; objects not in the pool are considered irrelevant. This method is regularly used in evaluation of text retrieval systems where evaluating all documents in the collection is practically infeasible.

A first observation is that SVCN outperforms WordNet and ARIA in all evaluations including the end-to-end evaluation. One exception is the result for recall in the downward strategy of the end-to-end evaluation; ARIA performs slightly better. The high scores of SVCN can be explained from its reasonably clean hierarchy and high similarity to the target vocabulary AAT. Evaluation of individual correspondences gives SVCN a precision of around 0.90 for all different weighting schemes. The different precision numbers lie around 0.70 for ARIA and WordNet. This suggests that a weakly structured, small vocabulary such as ARIA can be aligned with approximately the same precision as a large, richly structured vocabulary such as WordNet.

The variations in frequency of use of correspondences are most pronounced in WordNet. This can be explained from the fact that the proportion of WordNet concepts that has a correspondence to AAT is relatively small (13%). Queries for concepts without a correspondence to AAT will be reformulated to related concepts that do have a correspondence to AAT. This causes concepts that are central nodes in the hierarchy to get potentially high frequency counts. In line with this finding, correcting the most frequent correspondences gives a significantly higher precision than correcting randomly selected correspondences.

For ARIA, correcting frequent correspondences also showed a clear improvement of the results. This is not entirely expected, since ARIA has relatively many correspondences and ARIA's frequency distribution is less pronounced. The effect is partly due to the fact that the precision of the frequent stratum is lower than the precision of the infrequent stratum. The results suggest that for both WordNet and Aria, an evaluation that takes into account the frequency of use will result in a more realistic estimation of application performance than an evaluation that does not take this into account.

For SVCN, the frequency based weighting did not make a difference, nor did the correction of frequent correspondences. This lack of effect can be explained from two observations: (a) the precision of SVCN is already high and therefore correction will have less effect; and (b) the frequency distribution of SVCN correspondences is relatively gradual, so that the most frequent stratum has less influence than in the WordNet case. We conclude that in cases where only a small portion of a vocabulary can be aligned to a target vocabulary, for example when topical overlap is small, an estimation of the most frequently used correspondences gives a realistic image of application performance. In these cases it will be cost-effective to manually correct (only) the frequently used correspondences.

The comparison against a reference alignment produces a clear ordering of the three alignments, in both precision and recall: SVCN is best, followed by ARIA and finally WordNet. The alignment of WordNet has a low recall (0.45 and 0.47) compared to the other vocabularies. A possible cause is the size of WordNet and the relatively low number of correspondences that was found. Although we have no clear explanation, the effect is reflected in the end-to-end evaluation; WordNet has a remarkably low recall when using the downward strategy.

When comparing the 'traditional' precision and recall scores to those based on semantic distance, we see a clear difference between the two measures in

the results of ARIA (an average difference of 7%). A difference is notable for SVCN (average of 4%) although less clear, and almost no difference is visible for WordNet (1%). This is mirrored in the end-to-end evaluation, where the differences between a binary scale and a 6-point scale show the same trend: large differences for ARIA (an average of 13%), small differences for SVCN (6%) and no differences for WordNet (2%). An explanation is that ARIA returns many results that are only moderately relevant, while WordNet returns mainly highly relevant results. For applications in which users expect to see also moderately relevant results, an evaluation based on semantic distance better reflects the quality of the alignment.

We see two practical consequences of our analyses. First, we conclude that two vocabularies that show a stark resemblance to each other with respect to structure and topical overlap, can be aligned with such high precision and recall that manual creation or correction of the alignment has little added value. This holds in particular for vocabularies that share a common source, such as SVCN and AAT. Second, a vocabulary with a weak structure is no impediment for a high-quality alignment. The ontological flaws of ARIA did not result in a worse alignment than the reasonable structure of WordNet.

Acknowledgements

The authors would like to thank Willem van Hage, Alistair Vardy, Tom Moons, Niels Schreiber and the members of the E-Culture project. The authors were supported by the NWO projects: CHIME, CHOICE and STITCH and the TELplus project.

References

[1] B. Ashpole, M. Ehrig, J. Euzenat, and H. Stuckenschmidt, editors. *Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies*.

[2] W. P. Brink, van den and P. Koele. *Statistiek*, volume 3. Boom, Amsterdam, The Netherlands, 2002. ISBN 90 5352 705 2.

[3] A. Budanitsky and G. Hirst. Semantic distance in wordnet: an experimental application oriented evaluation of five measures. In *Proceedings of the NACCL 2001 Workshop on WordNet and other lexical resources*, pages 29–34, Pittsburgh, PA.

[4] M. Ehrig and J. Euzenat. Relaxed precision and recall for ontology matching. In Ashpole et al. [1].

[5] J. Euzenat. Semantic precision and recall for ontology alignment evaluation. In Manuela M. Veloso, editor, *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 348–353, 2007.

[6] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007. ISBN 3-540-49611-4.

[7] J. Euzenat, H. Stuckenschmidt, and M. Yatskevich. Introduction to the ontology alignment evaluation 2005. In Ashpole et al. [1].

[8] J. Euzenat, M. Mochol, P. Shvaiko, H. Stuckenschmidt, O. Šváb, V. Svátek, W.R. van Hage, and M. Yatskevich. Results of the OAEI 2006. In B. Ashpole, M. Ehrig, J. Euzenat, and H. Stuckenschmidt, editors, *Ontology Matching*, volume 225 of *CEUR Workshop Proceedings*, 2006.

[9] J. Euzenat, A. Isaac, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, O. Šváb, V. Svátek, W.R. van Hage, and M. Yatskevich. First results of the OAEI 2007. In B. Ashpole, M. Ehrig, J. Euzenat, and H. Stuckenschmidt, editors, *Ontology Matching*, CEUR Workshop Proc., 2007.

[10] Wei Hu and Yuzhong Qu. Discovering simple mappings between relational database schemas and ontologies. In *Proceedings of the International Semantic Web Conference*, pages 225 – 238, 2007.

[11] A. Isaac and S. Wang. Evaluation issues at the library testcase of oaei 2007. Accepted for publication in ESWC 2008.

[12] A. Isaac, C. Zinn, H. Matthezing, L. van der Meij, S. Schlobach, and S. Wang. The value of usage scenarios for thesaurus alignment in cultural heritage context. In *Proceedings of International Workshop on Cultural Heritage on the Semantic Web, ISWC2007*, Korea.

[13] Y. Kalfoglou and M. Schorlemmer. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18(1):1–31, 2003. ISSN 0269-8889.

[14] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), 2002. ISSN 1532-2882.

[15] C. Leacock and M. Chodorow. *Combining Local Context and WordNet Similarity for Word Sense Identification*, chapter 11, pages 265 – 285. MIT Press, 1998.

[16] T. Peterson. *Introduction to the Art and Architecture Thesaurus*. Oxford University Press, 1994.

[17] A. Th. Schreiber, A. Amin, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, L. Hollink, Z. Huang, J. van Kersen, M. de Niet, B. Omelayenko, J. van Ossenbruggen, R. Siebes, J. Taekema, J. Wielemaker, and B.J. Wielinga. Multimedian e-culture demonstrator. In *the Semantic Web Challenge at the Fifth International Semantic Web Conference*, Athens, GA, USA, November 2006.

[18] P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics*, 3730:146–171, 2005.

[19] M. van Assem, A. Gangemi, and G. Schreiber. RDF/OWL Representation of WordNet. W3C Working Draft, World Wide Web Consortium, June 2006.

[20] W.R. van Hage, A. Isaac, and Z. Aleksovski. Sample evaluation of ontology-matching systems. In *Proceedings of the Fifth International Evaluation of Ontologies and Ontology-based Tools*, Busan, Korea, 2007.