Sample Evaluation of Ontology-Matching Systems

Willem Robert van Hage[1,2], Antoine Isaac[1], and Zharko Aleksovski[3]

[1] Vrije Universiteit, Amsterdam
[2] TNO Science & Industry, Delft
[3] Philips Research, Eindhoven
{wrvhage,aisaac,zharko}@few.vu.nl

Abstract. Ontology matching exists to solve practical problems. Hence, methodologies to find and evaluate solutions for ontology matching should be centered on practical problems. In this paper we propose two statistically-founded evaluation techniques to assess ontology-matching performance that are based on the application of the alignment. Both are based on sampling. One examines the behavior of an alignment in use, the other examines the alignment itself. We show the assumptions underlying these techniques and describe their limitations.

1 Introduction

The advent of the Semantic Web has led to the development of an overwhelming number[4] of ontologies. Therefore, cross-referencing between these ontologies by means of ontology matching is now necessary. Ontology matching has thus been acknowledged as one of the most urgent problems for the community, and also as one of the most scientifically challenging tasks in semantic-web research.

Consequently, many matching tools have been proposed, which is a mixed blessing: *comparative* evaluation of these tools is now required to guide both ontology-matching research and application developers in search of a solution. One such effort, the Ontology Alignment Evaluation Initiative[5] (OAEI) provides a collaborative comparison of state-of-the-art mapping systems which has greatly accelerated the development of high-quality techniques. The focus of the OAEI has been mainly on comparing mapping techniques for research.

Good evaluation of ontology-matching systems takes into account the purpose of the alignment.[6] Every application has different requirements for a matching system. Some applications use rich ontologies, others use simple taxonomies. Some require equivalence correspondences, others subsumption or even very specific correspondences such as artist-style or gene-enzyme. Also, the scope of concepts and relations is often determined by unwritten application-specific rules (*cf.* [2]). For example, consider the subclass correspondence between the concepts Gold and Jewelry. This correspondence holds if the scope of Gold is limited to the domain of jewelry. Otherwise the two would just be related terms. In either case, application determines relevance.

---

[4] http://swoogle.umbc.edu indexes over 10,000 ontologies by 2007.

[5] http://oaei.ontologymatching.org

[6] In this paper we use the definitions as presented in [1]: An ontology matching system produces a set of correspondences called an alignment.

The best way to evaluate the quality of an alignment is trough extensive practical use in real-world applications. This, however, is usually not feasible. The main reason for this is usually lack of time (*i.e.* money). Benchmarks and experiments using synthesized ontologies can reveal the strengths and weaknesses of ontology-matching techniques, but disregard application-specific requirements. Therefore, the second best option is to perform an evaluation that mimics actual usage. Either by performing a number of typical usage scenarios or by specifying the requirements an application has for the alignment and then testing whether these requirements are met. The final measure for system performance in practice is user satisfaction. For the evaluation of matching systems, this means that a set of correspondences is good if users are satisfied with the effect the correspondences have in an application.

Most current matching evaluation metrics simulate user satisfaction by looking at a set of assessed correspondences. For example, Recall expresses how many of the assessed correspondences are found by a system. This has two major problems. (i) Some correspondences have a larger logical consequence than others. That is to say, some correspondences subsume many other correspondences, while some only subsume themselves. This problem is addressed quite extensively in [3] and [4]. (ii) Correct correspondences do not automatically imply happy users. The impact of a correspondence on system performance is determined not only by its logical consequence, but also by its relevance to the user's information need. A correspondence can be correct and have many logical implications, but be irrelevant to the reasoning that is required to satisfy the user. Also, some correspondences have more impact than others.

In the following sections we propose two alternative approaches to include relevance into matching evaluation, one based on *end-to-end evaluation* (Sec. 2) and one based on *alignment sample evaluation* (Sec. 3). Both approaches use sample evaluation, but both what is sampled and the sample selection criteria are different. The former method uses sample queries, disregarding the alignment itself, and hence providing objectivity. The latter uses sample sets of correspondences which are selected in such a way that they represent different requirements of the alignment. We investigate the limitations of these statistical techniques and the assumptions underlying them. Furthermore, we calculate upper bounds to the errors caused by the sampling. Finally, in Sec. 4 we will demonstrate the workings of the latter of the two evaluation methods in the context of the OAEI 2006 food track.

2   End-to-end Evaluation

This approach is completely system-performance driven, based on a sample set of representative information needs. The performance is determined for each trial information need, using a measure for user satisfaction. For example, such an information need could be *"I would like to read a good book about the history of steam engines."* and one could use *F*-score or the Mean-Reciprocal Rank[7] of the best book in the result list, or the time users spent to find an answer. The set of trials is selected such that it fairly represents different kinds of usage, *i.e.* more common cases receive more trials. Real-life topics should get adequate representation in the set of trials. In practice the trials

---

[7] One over the rank of the best possible result, *e.g.* 1/4 if the best result is the fourth in the list.

are best constructed from existing usage data, such as log files of a baseline system. Another option is to construct the trials in cooperation with domain experts. A concrete example of an end-to-end evaluation is described in [5]. In their paper, Voorhees and Tice explicitly describe the topic construction method and the measure of satisfaction they used for the end-to-end evaluation of the TREC-9 question-answering track. The size and construction methods of test sets for end-to-end retrieval have been investigated extensively in the context of information retrieval evaluation initiatives such as TREC [6], CLEF, and INEX[8]. When all typical kinds of usage are fairly represented in the sample set, the total system performance can be acquired by averaging the scores.[9] To evaluate the effect of an ontology alignment, one usually compares it to a baseline alignment in the context of the same information system. By changing the alignment while keeping all other factors the same, the only thing that influences the results is the alignment. The baseline alignment can be any alignment, but a sensible choice is a trivial alignment based only on simple lexical matching.

Comparative End-to-end Evaluation

n	number of test trials (*e.g.* information system queries) in the evaluation sample
A, B	two ontology-matching systems
A_i	outcome of the evaluation metric (*e.g.* Semantic precision [3]) for the i-th test trial for system A
$I[A_i > B_i] = \begin{cases} 1 & A_i > B_i \\ 0 & A_i \leq B_i \end{cases}$	interpretation function that tests outperformance
$S_+ = \sum I[A_i > B_i]$	number of trials for which system A outperforms system B

To compare end-to-end system performances we determine whether one system performs better over a significant number of trials. There are many tests for statistical significance that use pairwise comparisons. Each test can be used under different assumptions. A common assumption is the normal distribution of performance differences: small differences between the performance of two systems are more likely than large differences, and positive differences are equally likely as negative differences. However, this is not very probable in the context of comparative evaluation of matching systems. The performance differences between techniques are usually of a much greater magnitude than estimation errors. There are many techniques that improve performance on some queries while not hurting performance on other queries. This causes a skewed distribution of the performance differences. Therefore, the most reliable test is the Sign-test [8, 9]. This significance test only assumes that two systems with an equal performance are equally likely to outperform each other for any trial. It does not take

---

[8] respectively `http://trec.nist.gov`, `http://www.clef-campaign.org`, and `http://inex.is.informatik.uni-duisburg.de`

[9] A more reliable method for weighted combination of the scores that uses the variance of each performance measurement is described in [7].

into account how much better a system is, only in how many cases a system is better. The test gives reliable results for at least 25 trials. It needs relatively large differences to proclaim statistical significance, compared to other statistical tests. This means statistical significance calculated in this way is *very* strong evidence.

To perform the Sign-test on the results of systems A and B on a set of n trials, we compare their scores for each trial, $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$. Based on these outcomes we compute S_+, the total the number of times A has a better score than B. For example, the number of search queries for which A retrieves better documents than B. The null-hypothesis is that the performance of A is equal to that of B. This hypothesis can be rejected at a confidence level of 95%[†] if

$$\frac{2 \cdot S_+ - n}{\sqrt{n}} > 1.96$$

For example, in the case of 36 trials, system A performs significantly better than system B when it outperforms system B in at least 23 of the 36 trials.

3   Alignment Sample Evaluation

Another evaluation approach is to assess the alignment itself. However, in practice, it is often too costly to manually assess all the correspondences. A solution to this problem is to take a small *sample* from the whole set of correspondences [10]. This set is manually assessed and the results are generalized to estimate system performance on the whole set of correspondences. As opposed to the elegant abstract way of evaluating system behavior provided by *end-to-end evaluation*, *alignment sample evaluation* has many hidden pitfalls. In this section we will only investigate the caveats that are inherent to sample evaluation. We will not consider errors based on non-sampling factors such as judgement biases, peculiarities of the ontology-matching systems or ontologies, and other unforeseen sources of evaluation bias.

Simple Random Sampling

p true proportion of the samples produced that is correct (unknown)
n number of sample correspondences used to approximate p
$\hat{P}$ approximation of p based on a sample of size n
δ margin of error of $\hat{P}$ with 95% confidence

The most common way to deal with this problem is to take a small *simple random sample* from the whole set of correspondences. Assessing a set of correspondences can be seen as classifying the correspondences as *Correct* or *Incorrect*. We can see the output of a matching system as a *Bernoulli random variable* if we assign 1 to every *Correct* correspondence and 0 to each *Incorrect* correspondence it produces. The true

---

[†] About 95% of the cases fall within 1.96 times the standard deviation from the mean of the normal or binomial distribution. In the derivations we use 2 instead of 1.96 for the sake of simplicity. This guarantees a confidence level of more than 95%.
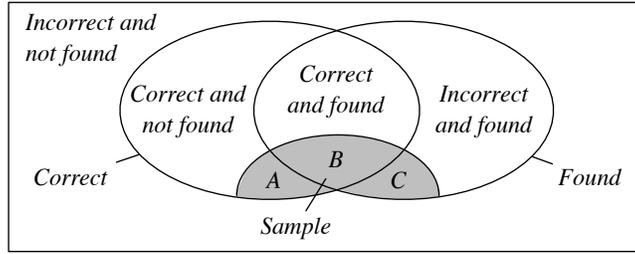
Fig. 1. Venn diagram to illustrate sample evaluation. $A \cup B$ is a sample of the population of Correct correspondences. $B \cup C$ is a sample of the population of Found correspondences.

mapping from ontology X to ontology Y mapping from ontology Y to ontology X
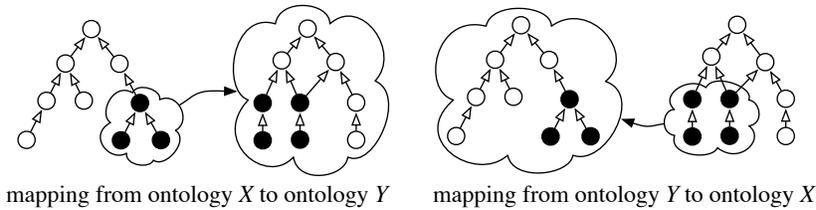
Fig. 2. Concepts to consider when creating a sample for Recall evaluation based on a topic. Black concepts are "on topic", white concepts "off topic". For example, the black concepts have something to do with steam engines and the white concepts do not. Concepts to consider for sample correspondences are marked by clouds. This avoids bias against cross-topic correspondences.

Precision of a system is the probability with which this random variable produces a 1, p. We can approximate this p by the proportion of 1's in a *simple random sample* of size n. With a confidence of 95% this approximation, $\hat{P}$, lies in the interval:

$$\hat{P} \in [p - \delta, p + \delta] \quad \text{where} \quad \delta = \frac{1}{\sqrt{n}} \tag{1}$$

Both Precision and Recall can be estimated using samples. In the case of Precision we take a random sample from the output of the matching system, *Found* in Fig. 1. In this figure the sample for Precision is illustrated as $B \cup C$. The results for this sample can be generalized to results for the set of all *Found* correspondences. In the case of Recall we take a random sample from the set of all correct correspondences, *Correct* in Fig. 1. The sample for Recall is illustrated as $A \cup B$. The results for this sample can be generalized to results for the set of all *Correct* correspondences.

A problem with taking a random sample from all *Correct* correspondences is it is unknown which correspondences are correct and which are incorrect a priori. A proper random sample can be taken by randomly selecting correspondences between all possible correspondences between concepts from the two aligned ontologies, *i.e.* a subset of the cartesian product of the sets of concepts from both ontologies. Each correspondence has to be judged to filter out all incorrect correspondences. This can be very time-consuming if there are relatively few valid correspondences in the cartesian product. The construction time of the sample of correct correspondences can be reduced

by only judging parts of the ontologies that have a high topical overlap. For example, one can only consider all correct mappings between concepts having to do with steam engines. (*cf. e.g.* [11]) It is important to always match concepts about a certain topic in ontology X to *all* concepts in ontology Y, and all concepts about the same topic in ontology Y to *all* concepts in ontology X. This is illustrated in Fig. 2. This avoids a bias against correspondences to concepts outside the sample topic.

There are two caveats when applying this approximation method. (i) A sample of correct mappings constructed in this way is arbitrary, but not completely random. Correspondences in the semantic vicinity of other correspondences have a higher probability of being selected than "loners". This means ontology matching techniques that employ structural aspects of the ontologies are slightly advantaged in the evaluation. (ii) The method works under the assumption that correspondences inside a topic are equally hard to derive as correspondences across topics.

Stratified Random Sampling

N size of the entire population, *e.g.* the set of all correct correspondences

h one stratum of the entire population

N_h size of stratum h

n_h number of sample correspondences used to approximate p of stratum h

$\hat{P}_h$ approximation of p for the correspondences in stratum h

A better way than *simple random sampling* to perform sample evaluation is *stratified random sampling*. In stratified sampling, the population (*i.e.* the entire set of correspondences used in the evaluation) is first divided into subpopulations, called *strata*. These strata are selected in such a way that they represent parts of the population with a common property. Useful distinctions to make when stratifying a set of correspondences are: different alignment relations (*e.g.* equivalence, subsumption), correspondences in different domains (*e.g* cats, automobiles), different expected performance of the matching system (*e.g.* hard and easy parts of the alignment), or different levels of importance to the use case (*e.g* mission critical versus nice-to-have). The strata form a partition of the entire population, so that every correspondence has a non-zero probability to end up in a sample. Then a sample is drawn from each stratum by *simple random sampling*. These samples are assessed and used to score each stratum, treating the stratum as if it were an entire population. The approximated proportion and margin of error can be calculated with *simple random sampling*.

Stratified random sampling for the evaluation of alignments has two major advantages over simple random sampling. (i) The separate evaluation of subpopulations makes it easier to investigate the conditions for the behavior of matching techniques. If the strata are chosen in such a way that they distinguish between different usages of the correspondences, we can draw conclusions about the behavior of the correspondences in a use case. For example, if a certain matching technique works very well on chemical concepts, but not on anatomical concepts, then this will only come up if this division is made through stratification. (ii) Evaluation results for the entire population acquired by combining the results from stratified random sampling are more precise than those of

simple random sampling. With simple random sampling there is always a chance that the sample is coincidentally biased against an important property. While every property that is distinguished in the stratification process will be represented in the sample.

The results of all the strata can be combined to one result for the entire population by weighing the results by the relative sizes of the strata. Let N be the size of the entire population and $N_1, \ldots, N_L$ the sizes of strata 1 to L, so that $N_1 + \cdots + N_L = N$. Then the weight of stratum h is N_h/N. Let n_h be the size of the *simple random sample* in stratum h and $\hat{P}_h$ be the approximation of proportion p in stratum h by the sample of size n_h. We do not require the sample sizes $n_1, \ldots, n_L$ to be equal, or proportional to the size of the stratum. The approximated proportion in the entire population, $\hat{P}$, can be calculated from the approximated proportions of the strata, $\hat{P}_h$, as follows:

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

Due to the fact that the variance of the binomial distribution is greatest at $p = 0.5$, we know that the greatest margin-of-error occurs when $\hat{P} = 0.5$. That means that with a confidence of 95% the approximation of $\hat{P}$ lies in the interval:

$$\hat{P} \in [p - \delta, p + \delta] \quad \text{where} \quad \delta = \frac{1}{\sqrt{N}} \sqrt{\sum_{h=1}^{L} (\frac{N_h}{n_h} - 1)} \tag{2}$$

Comparative Alignment Sample Evaluation

p_A true proportion of the correspondences produced by system A that is correct (unknown)

$\hat{P}_A$ sample approximation of p_A

$\hat{P}_{A,h}$ $\hat{P}_A$ in stratum h

To compare the performance of two systems, A and B, using sample evaluation, we calculate their respective $\hat{P}_A$ and $\hat{P}_B$ and check if their margins of error overlap. If this is not the case, we can assume with a certain confidence that p_A and p_B are different, and hence that one system is significantly better than the other. For *simple random sampling* this can be calculated as follows:

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\frac{\hat{P}_A(1 - \hat{P}_A)}{n} + \frac{\hat{P}_B(1 - \hat{P}_B)}{n}} \tag{3}$$

For *stratified random sampling* this can be calculated as follows:

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\sum_{h=1}^{L} \frac{\hat{P}_{A,h}(1 - \hat{P}_{A,h})}{N} \left(\frac{N_h}{n_h} - 1\right) + \sum_{h=1}^{L} \frac{\hat{P}_{B,h}(1 - \hat{P}_{B,h})}{N} \left(\frac{N_h}{n_h} - 1\right)} \tag{4}$$

For both methods the maximum difference needed to distinguish P_A from P_B with a confidence of 95% is $2/\sqrt{2n}$. So if, depending on the type of sampling performed, equation (3) or (4) holds, there is a significant difference between the performance of system A and B.

4 Alignment Sample Evaluation in Practice

In this section we will demonstrate the effects of *alignment sample evaluation* in practice by applying *stratified random sampling* on the results of the OAEI 2006 food track[10] for the estimation of Precision and we will calculate the margin of error caused by the sampling process.

The OAEI 2006 food track is a thesaurus matching task between the Food and Agriculture Organisation of the United Nations (FAO) AGROVOC thesaurus and the thesaurus of the United States Department of Agriculture (USDA) National Agricultural Library (NAL). Both thesauri are supplied to participants in SKOS and OWL Lite[11]. The alignment had to be formulated in SKOS Mapping Vocabulary[12] and submitted in the common format for alignments[13]. A detailed description of the OAEI 2006 food track can be found in [12, 13].

Five teams submitted an alignment: Falcon-AO, COMA++, HMatch, PRIOR, and RiMOM. Each alignment consisted only of one-to-one semantic equivalence correspondences. The size of the five alignments is shown below.

system	RiMOM	Falcon-AO	Prior	COMA++	HMatch	all systems
# Found	13,975	13,009	11,511	15,496	20,001	31,112

The number of unique *Found* correspondences was 31,112. The number of *Correct* correspondences can be estimated in the same order of magnitude. In our experience, voluntary judges can only reliably assess a few hundred correspondences per day. That means this means assessing all the *Found* correspondences in the alignments would already take many judges a few weeks of full-time work. This is only feasible with significant funding. Thus, we performed a sample evaluation.

During a preliminary analysis of the results we noticed that the performance of the different systems was quite consistent for most topics, except correspondences between taxonomical concepts (*i.e.* names of living organisms such as "Bos Taurus") with latin names where some systems performed noticeably worse than others. This was very surprising given that there was a straightforward rule to decide the validity of a taxonomical correspondence, due to similar editorial guidelines for taxonomical concepts in the two thesauri. Two concepts with the same preferred label and some ancestors with the same preferred label are equivalent. Also, when the preferred label of one concept is literally the same as the alternative label of the other and some of their ancestors have the same preferred label they are equivalent. For example, the African elephant in AGROVOC has a preferred label "African elephant" and an alternative label "Loxodonta africana". In NALT it is the other way around.

These rules allowed us to semi-automatically assess the taxonomical correspondences. This was not possible for the other correspondences. So we decided to separately evaluate correspondences from and to taxonomical concepts. We also noticed

---

[10] http://www.few.vu.nl/~wrvhage/oaei2006

[11] The conversion from SKOS to OWL Lite was provided by Wei Hu.

[12] http://www.w3.org/2004/02/skos/mapping/spec

[13] http://oaei.ontologymatching.org/2006/align.html

that most other correspondences were very easy to judge, except correspondences between biochemical concepts (*e.g.* "protein kinases") and substance names (*e.g.* "tryptophan 2,3-dioxygenase"). These required more than a layman's knowledge of biology or chemistry. So we decided to also evaluate biological and chemical concepts separately, with different judges. This led to three strata: taxonomical correspondences, biological and chemical correspondences, and the remaining correspondences. The sizes of the strata, along with the size of the evaluated part of the stratum and the corresponding stratum weights are shown below.

stratum topic	stratum size (N_h)	sample size (n_h)	stratum weight (N_h/N)
taxonomical	18,399	18,399	0.59
biological and chemical	2,403	250	0.08
miscellaneous	10,310	650	0.33
all strata	31,112	21,452	

Precision estimates using these strata have a maximum margin of error of:

$$2 \cdot \sqrt{\frac{0.5 \cdot (1 - 0.5)}{31112} \cdot \left(\left(\frac{18399}{18399} - 1\right) + \left(\frac{2403}{250} - 1\right) + \left(\frac{10310}{650} - 1\right) \right) \cdot 2} \approx 3.8\%$$

at a confidence level of 95%. That means that, under the assumption that there are no further biases in the experiment, a system with 82% Precision outperforms a system with 78% Precision with more than 95% confidence.

If, for example, we are interested in the performance of a system for the alignment of biological and chemical concepts and use the sample of 250 correspondences to derive the performance on the entire set of 2,403 correspondences our margin of error would be $1/\sqrt{250} \approx 6.3\%$. Comparison of two systems based on only these 250 sample biological and chemical correspondences gives results with a margin of error of $2/\sqrt{2 \cdot 250} \approx 8.9\%$. That means with a confidence level of 95% we can distinguish a system with 50% Precision from a system with 59% Precision, but not from a system with 55% Precision.

5 Conclusion

We presented two alternative techniques for the evaluation of ontology-matching systems and showed the margin of error that comes with these techniques. We also showed how they can be applied and what the statistical results mean in practice in the context of the OAEI 2006. Both techniques allow a more application-centered evaluation approach than current practice.

Apart from sampling errors we investigated in this paper, there are many other possible types of errors that can occur in an evaluation setting. (Some of which are discussed in [14].) Other sources of errors remain a subject for future work. Also, this paper leaves open the question of which technique to choose for a certain evaluation effort. For example, when you want to apply evaluation to find the best ontology matching system for a certain application. The right choice depends on which technique is more cost effective. In practice, there is a trade-off between cheap and reliable evaluation: With limited resources there is no such thing as absolute reliability. Yet, all the questions we

have about the behavior of matching systems will have to be answered with the available evaluation results. The nature of the use case for which the evaluation is performed determines which of the two approaches is more cost effective. Depending on the nature of the final application, evaluation of end-to-end performance will sometimes turn out to be more cost effective than investigating the alignment, and sometimes the latter option will be a better choice. We will apply the techniques presented in this paper to the food, environment, and library tasks of the forthcoming OAEI 2007.[14] This should give us the opportunity to further study this subject.

Acknowledgments

References

1. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag, Heidelberg (DE) (2007)
2. Šváb, O., Svátek, V., Stuckenschmidt, H.: A study in empirical and casuistic analysis of ontology mapping results. In: Proc. of the European Semantic Web Conf. (ESWC). (2007)
3. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: Proc. of IJCAI 2007. (2007) 348–353
4. Ehrig, M., Euzenat, J.: Relaxed precision and recall for ontology matching. In: Proc. of K-CAP 2005 workshop on integrating ontologies. (2005) 25–32
5. Voorhees, E., Tice, D.: Building a question answering test collection. In: Proc. of SIGIR. (2000)
6. Voorhees, E.: Variations in relevance judgments and the measurement of retrieval effectiveness. In: Research and Development in Information Retrieval. (1998) 315–323
7. Meier, P.: Variance of a weighted mean. Biometrics **9**(1) (1953) 59–73
8. Hull, D.: Evaluating evaluation measure stability. In: Proc. of SIGIR 2000. (2000)
9. van Rijsbergen, C.J.: Information Retrieval. Butterworths (1979)
10. Cochran, W.G.: Sampling Techniques. John Wiley & Sons, Inc. (1977)
11. Wang, S., Isaac, A., van der Meij, L., Schlobach, S.: Multi-concept alignment and evaluation. In: Proc. of the Int. Workshop on Ontology Matching. (2007)
12. Euzenat, J., Mochol, M., Shvaiko, P., Stuckenschmidt, H., Šváb, O., Svátek, V., van Hage, W.R., Yatskevich, M.: Results of the ontology alignment evaluation initiative (2006)
13. Shvaiko, P., Euzenat, J., Stuckenschmidt, H., Mochol, M., Giunchiglia, F., Yatskevich, M., Avesani, P., van Hage, W.R., Šváb, O., Svátek, V.: Description of alignment evaluation and benchmarking results. KnowledgeWeb Project deliverable D2.2.9 (2007)
14. Avesani, P., Giunchiglia, F., Yatskevich, M.: A large scale taxonomy mapping evaluation. In: Proc. of the Int. Semantic Web Conf. (ISWC). (2005)

---

[14] http://oaei.ontologymatching.org/2007/

[15] http://www.vl-e.nl