UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

BEST PRACTICES FOR ONTOLOGY MATCHING TOOLS EVALUATION

Aliaksandr Autayeu, Vicenzo Maltese, and Pierre Andrews

August 2009

Technical Report # DISI-09-046

Also: Poster at the ISWC Ontology Matching Workshop (OM 2009), 25th October 2009, Washington DC, USA

Best Practices for Ontology Matching Tools Evaluation

Aliaksandr Autayeu, Vicenzo Maltese, and Pierre Andrews

Dipartimento di Ingegneria e Scienza dell'Informazione
Università degli Studi di Trento, Trento, Italy

Abstract. In the current state of the art in ontology matching, diverse golden standards are used to evaluate the algorithms. In this paper we show that by following appropriate rules in their construction and use, the quality of the evaluations can be significantly improved, particularly in the accuracy of the precision and recall measures obtained.

1 Introduction

In the recent years, as a valid solution to the semantic heterogeneity problem, several matching tools have been proposed. These tools identify nodes in two schemas (e.g., database and XML schemas, classifications, thesauri or ontologies), that are syntactically or semantically related. We focus on matching techniques, which find semantic correspondences between nodes [1, 2], such as SMatch [3–5].

In this paper we discuss issues related to evaluating tools that output rich mappings. In particular, we consider the types of output similar to the one provided by SMatch. In SMatch, the two schemas in input are preliminary converted into lightweight ontologies [6]. With the conversion, each natural language node label is translated into a propositional Description Logic (DL) formula encoding the meaning of the node. The algorithm then computes the *mapping*, also called alignment, containing the set of semantic correspondences between nodes in the two ontologies. Each correspondence is given a semantic relation in the set: *disjointness* ($\perp$), *equivalence* ($\equiv$), *less general* ($\sqsubseteq$) and *more general* ($\sqsupseteq$).

Most of the tools for semantic matching identify only equivalence, some of them identify less and more general correspondences, but only a few of them include explicit disjointness [7]. Reflecting this, an overwhelming majority of available golden standards are targeted at evaluating mappings containing only equivalence correspondences. In this paper we explain why evaluating a rich mapping using such a golden standard makes results imprecise and provide best practices to make results of evaluation and comparison between different tools fair and more accurate.

The rest of the paper is organized as follows. We first introduce the related work in the field of ontology matching and specifically of their evaluation. Section 3 discusses the issues that arise when computing precision and recall

measures on large datasets. Section 4 introduces the notion of redundant correspondences in a mapping and discusses their influences on precision and recall calculation. Section 5 provides an evaluation that emphasizes the repercussions of evaluating with and without redundancy. Finally, Section 6 provides a set of conclusions on all of these issues and an outline of the best practices for the evaluation of ontology matching algorithms.

2 Related Work

The main focus of the ontology matching field are the matching techniques themselves (see [1] for a recent survey), in particular within the successful OAEI[1] campaign and rarely the evaluation methodology. We believe that the community can benefit from a greater attention to the evaluation issues raised by the use and the quality of existing golden standards.

The evaluation methodology of the ontology matching tools has been scarcely examined in the literature. For instance, [8, 9] report on general evaluation experiments and [10, 11] on domain-specific evaluation experiments, but they do not discuss any of the existing issues in the existing evaluation methodology.

Considerable attention has been paid to appropriateness and quality of the measures, such as standard precision and recall. For example, in [12] the authors propose a framework for generalizing precision and recall. The improvement of precision and recall measures continued in [13], where the authors propose semantic precision and recall. Later, these improvements were analyzed and further advanced in [14], where adaptations of the relaxed and semantic precision and recall to the normalized mapping are proposed.

Attention has also been brought to the mapping itself, such as in [15] where the authors propose to complement the precision and recall with new measures to take into account possible mapping incoherence, thus addressing the issues of internal logical problems of the mapping and the lack of reference mappings. In [16] two evaluation techniques are proposed. The first is practice-oriented and evaluates the behaviour of the mapping in use. The second focuses on the manual evaluation of a mapping sample and the generalization of the results.

Closer to our work the authors of [7] raise the issue of evaluating non-equivalence correspondences, pointing out that more and more systems start to produce correspondences with relations such as subsumption and disjointness. In particular, they discuss the issue of evaluating a mapping that contains redundant correspondences, that is, correspondences that can be logically derived from the others in the mapping. They compute precision both for the original set and the set from which the redundant correspondences are eliminated. We extend and correct their conclusions.

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

3 Computing Precision and Recall

Golden standards are fundamental for the computation of precision and recall, as they are normally used for the evaluation of ontology matching techniques [14]. They are typically manually constructed. A positive golden standard, that we denote with GS^+, contains correspondences which are considered correct by a human editor. On the other hand, a negative golden standard, that we denote with GS^-, contains correspondences which are considered wrong. In the ideal case, GS^- is the complement of GS^+. However, in large datasets it is practically impossible to annotate all of the possible correspondences (see Sect. 3.1) and thus, the golden standard is composed of three sets:

***GS*$^+$** the set of annotated node pairs that hold as true in the alignment (i.e. the correspondences that would be correct if they were discovered by the matching algorithm);

***GS*$^-$** the set of annotated node pairs that do not hold in the alignment (i.e. the correspondences that would be wrong if they were discovered by the matching algorithm);

Unk the set of node pairs for which there are no annotations (i.e. the relation between the two nodes is unknown).

If we denote the result of the matcher (i.e. the mapping) with Res, precision and recall can be computed as follows [17]:

$$Precision = \frac{TP}{TP + FP} = \frac{|Res \cap GS^+|}{|Res \cap GS^+| + |Res \cap GS^-|} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} = \frac{|Res \cap GS^+|}{|GS^+|} \tag{2}$$

In case GS^- is not available, precision can be approximated as follows:

$$Precision = \frac{|Res \cap GS^+|}{|GS^+|} \tag{3}$$

Where:

TP is the set of correspondences found by the algorithm that hold (True Positives),

FP is the set of correspondences found by the algorithm that do not hold (False Positives),

FN is the set of correspondences that the algorithm did not find (False Negatives).

These sets are illustrated in Fig. 1. For example, if for sake of simplicity we use numbers to indicate the correspondences, we could have:

$$Res = \{1, 2, 3, 4\} \quad GS^+ = \{1, 2, 5, 7, 9, 10\} \quad GS^- = \{3, 4, 6, 8\} \quad Unk = \{\}$$
$$Precision = \frac{2}{(2 + 2)} = 0.5 \qquad Recall = \tfrac{2}{6} = 0.33 \tag{4}$$

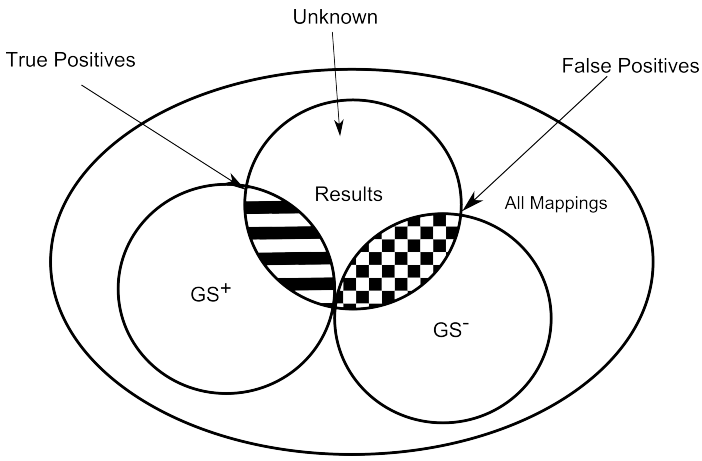


Fig. 1. True Positives, False Positives and Golden Standards

The precision gives an indication of the amount of noise that is retrieved by the matching algorithm (i.e. how many correct correspondences it returns) while the recall is a measure of the coverage of the algorithm (i.e. how many correspondences the algorithm found and missed).

In the next sections we first show how golden standards of different coverage or with different relations may lead to different evaluation results, and then we propose the best approach to follow for evaluation.

3.1 Coverage of the Golden Standard

Given two ontologies, respectively of size n and m, the size of a mapping, and therefore of both the negative and positive golden standards, can range from zero up to $n \times m$. In principle, in order to precisely construct them, one should inspect all possible $n \times m$ combinations of nodes in the two ontologies and consider all possible semantic relations which can hold between them. Constructing a golden standard in this way would allow a very precise computation of precision and recall. However, for large ontologies it would be costly or practically impossible. Our assumption is that this is the main reason why only a few golden standards are available and evaluation campaigns, such as the OAEI [18], tend to use very small ontologies. However, by using a small dataset for evaluation and comparison, there is a loss of statistical significance of the results and a dataset bias can be introduced towards one algorithm or the other.

When setting up exhaustive GS^+ and GS^- is not possible, the common practice is to compute them on a subset of the $n \times m$ pairs of nodes [19, 17]. At the beginning of this section we have denoted the set of non-inspected node pairs with Unk (unknown). The partial coverage clearly leads to an approximated evaluation. In particular, we cannot say anything about the subset $Res \cap Unk$ of correspondences identified by the matcher. However, if GS^+ and GS^- are

sampled properly, the precision and recall can still be evaluated in a statistically significant manner. For example, we could have reduced coverage:

$$Res = \{1, 2, 3, 4\} \quad GS^+ = \{1, 2, 7\} \quad GS^- = \{3, 6, 8\} \quad Unk = \{4, 5, 9, 10\}$$

$$Precision = \frac{2}{(2+1)} = 0.66 \quad Recall = \frac{2}{3} = 0.66 \quad Res \cap Unk = \{4\}$$

As it can be noted from this example, the results of such evaluations may be different from the real values (compare them with (4)). The relevance of the results depends on:

- the portion of the pairs covered;
- the ratio between the size of the GS^+ and the GS^-;
- their quality (as illustrated in the Sect. 4).

3.2 Comparing Semantic Relations

In the current state of the art, the available tools produce different kinds of relations in output. While most of the matching tools, such as Similarity Flooding [20], Cupid [21] and COMA [22] only produce *equivalence* relations, some tools, such as CSR [23] and AROMA [24] also produce *less general* and *more general* relations. At the best of our knowledge only ctxMatch [25], SMatch [3–5], MinSMatch [26] and Spider [7] produce explicit *disjointness* between nodes.

To compare tools producing different semantic relations in output, some criteria have to be established for a comparison to be correct. Different algorithms are usually compared without distinguishing among the different semantic relations produced and only the presence or absence of a relation between a pair of nodes is evaluated. This means, for instance, that $\sqsupseteq$, $\sqsubseteq$ and $\equiv$ are considered as the same. This approach can be used to compare heterogeneous correspondences, but provides imprecise results as it is unclear how the algorithms behave on each type of relation.

A different discourse has to be made for *disjointness* relations. Typically disjointness can be seen as a negative result, namely a clear indication of two completely unrelated nodes. Thus, the majority of matching tools do not consider them interesting to the users. As a consequence they do not compute them at all, but corresponding node pairs are rather put in the GS^-.

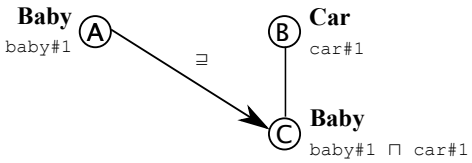


Fig. 2. Overlap between nodes A and B. Natural language labels are in bold with a corresponding DL formulas under them.

Consider the example in Fig. 2. The mapping element $< A, C, \sqsupseteq >$ is a correct result and as such should be part of the GS^+. In fact, given the semantics of lightweight ontologies, the meaning of the node C includes the meaning of the node B above it. What about the relation between A and B? They are not disjoint as they share C. The relation is rather an *overlap* (namely $A \cap B \neq \emptyset$). Discriminating the two cases above is fundamental both to conclude the right relations between the nodes and to correctly evaluate precision and recall of disjointness relations when they are explicitly computed by the matching tool. In fact, the main problem is that negative golden standards (when available) typically contain undifferentiated correspondences. For instance, the authors of [17] make no difference between *disjointness* and *overlap* relations.

3.3 Best Practice

When computing precision and recall, if the golden standard is small, some questions can be raised on the relevance of the results. In particular, it is unclear if the algorithm is biased toward the dataset and if the algorithm is portable. By constructing a large golden standard, the diversity of the correspondences that it contains is greater and thus the results of evaluations will be more statistically significant.

> *First recommendation.* Evaluations should be performed on large golden standards.

However, constructing a large golden standard is often difficult. To overcome this problem, in [17] the authors propose a semi-automatic procedure to construct a golden standard sampling. If the sampling is constructed fairly, the results can be considered as significant. However, in this case, the measures of precision and recall must take into account that several relations remain unknown (see Sect. 3.1).

> *Second recommendation.* To provide a good approximation of the precision and recall measures, the sampled golden standard must also provide a set of negative correspondences (that we call GS^-) in addition to the usual set of positive results. Both the positive and negative samples should cover a significant portion of the possible node pairs to be statistically significant.

In the current state of the art there is a large diversity of results returned by the matching algorithms. Some of them only provide information about the existence of a correspondence (which is typically interpreted as *equivalence*), while others specify the kind of semantic relation as either *more general*, *less general* or *equivalence*. Some algorithms even provide information about logical *disjointness* between nodes. Conducting an accurate comparison based on the different relations produced is clearly not possible (see Sect. 3.2).

Third recommendation. When comparing algorithms that produce different kinds of relations it is fundamental to specify how the analysis is performed and whether it takes into account the semantic relation type.

In addition, we want to stress the point that more research is required to build a methodology for evaluating algorithms producing *disjointness* relations. To the best of our knowledge, no evaluations take *disjointness* relations into account when measuring precision and recall.

4 Maximized and Minimized Golden Standards

In this section we use the notion of minimal mapping [26] to judge of the quality of a golden standard. The basic idea is that among all possible correspondences between two ontologies there are some redundant ones, which can be logically inferred from the other correspondences. Therefore, the minimal mapping is defined as the minimal set of (non-redundant) correspondences such that all the other (redundant) correspondences can be logically inferred from the non-redundant ones.

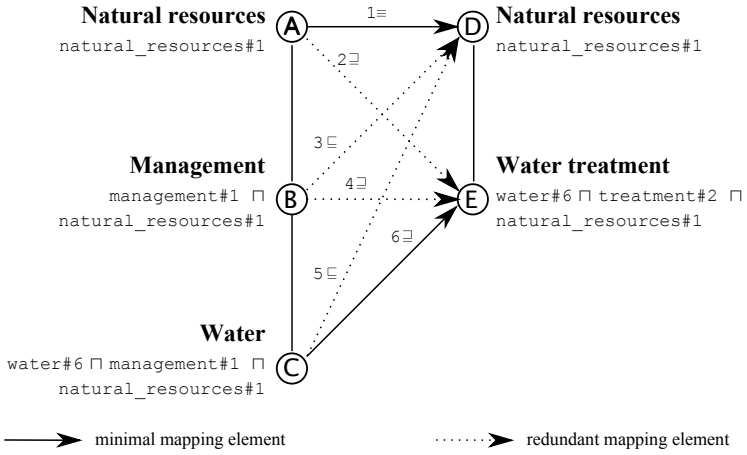


Fig. 3. The mapping between two lightweight ontologies. Original natural language labels are in bold.

Consider the example in Fig. 3 taken from [27]. It shows the minimal mapping (the solid arrows) and the mapping of maximum size (including the maximum number of redundant correspondences represented as dashed arrows) computed between two lightweight ontologies.

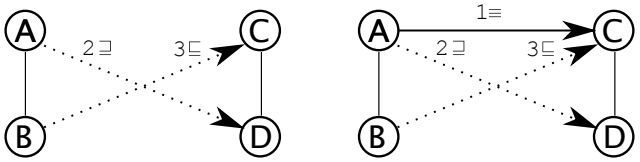
Using the notions briefly described above, we compute the following functions:

Min(mapping) removes the redundant correspondences from the mapping. We call the output of this function the *minimized mapping*;

Max(mapping) extends the mapping by computing all the correspondences which are redundant. We call the output of this function the *maximized mapping*.

The result of the Min and Max functions applied to a mapping are what [14] calls the "semantic reduction" and the "semantic closure", respectively. However, in the approach proposed in [26] (by restricting the application to lightweight ontologies) there are at least two fundamental advantages: (a) we can compute the redundant correspondences in time linear to the size of the two ontologies; (b) the set of all possible consequences in the maximized set is always finite and therefore precision and recall can always be computed.

In contrast with [7], we show that comparing the minimized versions of the mapping and the golden standards is not informative. Consider the examples in the Fig. 4.



(a) GS^+ **(b)** *Res*

Fig. 4. Minimization changing precision and recall

Suppose that all the displayed correspondences are correct. Notice that 2 and 3 follow from 1. Suppose that our golden standard (Fig. 4a), as it often happens with large datasets, is incomplete (contains only the correspondences 2 and 3) and thus we use precision formula (3). Suppose the matcher, being good enough, finds all displayed correspondences (Fig. 4b). In this case, the precision and recall figures are as follows:

$$GS^+ = \{2,3\} \quad Res = \{1,2,3\} \quad Precision = 0.66 \quad Recall = 1 \tag{5}$$

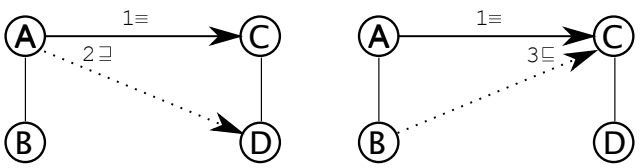
$$Min(GS^+) = \{2,3\} \quad Min(Res) = \{1\} \quad Precision = 0 \quad Recall = 0 \tag{6}$$

Compare (5) with (6) that shows the situation when minimized sets are used to calculate precision and recall figures. From this example we see that precision and recall figures are far from the real values and therefore unreliable.

Consider now the example in Fig. 5. The precision and recall figures are given in (7) for the original sets and in (8) for the maximized ones.

$$GS^+ = \{1,2\} \quad Res = \{1,3\} \quad Precision = 0.5 \quad Recall = 0.5 \tag{7}$$

$$Max(GS^+) = \{1,2,3\} \;\; Max(Res) = \{1,2,3\} \;\; Precision = 1 \;\; Recall = 1 \tag{8}$$



(a) GS^+ **(b)** *Res*

Fig. 5. Maximization changing precision and recall

In the maximized case precision and recall figures adhere to the real situation. Using maximized sets gives no preference to redundant or non-redundant correspondences and leads to more accurate results. In particular, recall figure better shows the amount of information actually found by the system. If we maximize the sets we also decrease the number of unknown correspondences and therefore we obtain a more accurate result. On the other hand, comparing the minimized versions of the mapping produced by the matcher and the GS^+ is not informative.

Maximizing a golden standard can also reveal some unexpected problems and inconsistencies. For instance, we can discover that even if GS^+ and GS^- are disjoint, $Max(GS^+)$ and $Max(GS^-)$ are not, and there is an overlap in their intersection. During our experiments with the TaxME2 golden standard [17], we have discovered that there are two correspondences in the intersection of GS^+ and GS^- and 2187 in the intersection of their maximized versions. These two correspondences are:

```
A: Top/Arts/Music/Styles/Jazz/Swing
B: Top/Entertainment/Music/Artists/By_Genre/Jazz/Swing

C: Top/Arts/Music/Styles/Polka
D: Top/Entertainment/Music/Artists/By_Genre/Folk_and_Traditional/Polka
```

In fact, in the GS^+ we found $A \sqsubseteq B$ and $C \equiv D$, while in the GS^- we found that $A \equiv B$ and $C \equiv D$.

To summarize:

> *Fourh recommendation.* When measuring precision and recall both the golden standards and the result of the matching algorithm should be maximized in order to contain all possible redundant correspondences.

5 Evaluation

We conducted several experiments to study the differences between precision and recall measures when comparing the minimized and maximized versions of the golden standards with the minimized and maximized versions of the mapping returned by SMatch.

Table 1. Golden standards used in experiments

Dataset pair	Node count	Max depth	ABF	Golden standards	
101/304	33/42	3/4	4.12/3.50	29/0/0/0	29^{+}
Topia/Icon	542/999	2/9	8.19/3.66	41/146/1166/3	1356^{+}
Source/Target	2857/6628	11/15	2.04/1.94	1096/990/179/0	2265^{+}
				1639/641/94/0	2374^{-}

Table 1 shows some indicators of the complexity of the golden standards we used in our experiments. The first two datasets come from OAEI. They describe publications and are called 101 and 304, respectively. The second two come from the arts domain and are referred to as Topia and Icon, respectively. The third two datasets have been extracted from the Looksmart, Google and Yahoo! directories and are referred to as Source and Target. The ABF column contains the average branching factor. The golden standards column contains details about the relations contained in the golden standards. The first four figures indicate the amount of $\equiv$ / $\sqsubseteq$ / $\sqsupseteq$ / $\perp$ correspondences in the alignment, respectively. The fifth figure indicates the total amount and kind of alignment: positive (GS^{+}), or negative (GS^{-}). For the first two datasets the golden standard is exhaustive and comes from OAEI. For the second two datasets the golden standard is crafted by experts manually. The third golden standard is described in [17].

Table 2. Precision and Recall for minimized, normal, and maximized sets

Dataset pair	Precision, %			Recall, %		
	min	res	max	min	res	max
101/304	32.47	9.75	69.67	86.21	93.10	92.79
Topia/Icon	16.87	4.86	45.42	10.73	20.00	42.11
Source/Target	74.88	52.03	48.40	10.35	40.74	53.30

Table 2 contains precision and recall figures calculated using standard precision and recall formulas (1) and (2). For the cases where no GS^{-} is provided, (3) is used instead of (1). In particular, these figures are the result of the comparison of the minimized mapping with the minimized golden standards (min), the original mapping with the original golden standards (res) and the maximized mapping with the maximized golden standards (max) respectively. As it can be noted from the measures obtained comparing the maximized versions with the

original versions, the performance of the algorithm is on average better than expected.

6 Conclusion

In this paper, we discussed a number of issues in evaluating ontology matching algorithms. These tools are used to find semantic correspondences between the nodes of two different ontologies in input. In the current state of the art, the performance of these algorithms is evaluated through the use of golden standards providing the set of correct correspondences between the ontologies. These golden standards can thus be used to measure precision and recall.

We proposed a set of best practices to follow when building such golden standards and to effectively use the golden standards to evaluate matching algorithms. By following these recommendations, the comparison of different matching algorithms will be more accurate. Here is the summary of the best practices we propose:

Golden Standard Construction The size of the golden standard must be as large as practically possible.

False Negative Sample The sampling must include a set of negative results to evaluate the number of false negatives returned by the algorithm.

Matching Element Types It is important to take into account the type of semantic relations used when comparing algorithms.

Redundancy We recommend to maximize both the golden standard and the result set to contain *all* redundant links before computing precision and recall.

We also discussed the issue of evaluating mappings including disjointness. In the current state of the art, even if some algorithms are able to identify such relations, no golden standard is available yet that explicitly provides true disjointness correspondences. It is thus currently impossible to compare the performance of such algorithms.

References

1. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. JoDS **4** (2005) 146–171
2. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer-Verlag (2007)
3. Giunchiglia, F., Yatskevich, M., Giunchiglia, E.: Efficient semantic matching. In: Proc. of EWSC. (2005) 272–289
4. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Semantic schema matching. In: Proc. of CoopIS. (2005) 347–365
5. Giunchiglia, F., Yatskevich, M.: Element level semantic matching. In: Proc. of Meaning Coordination and Negotiation workshop at ISWC. (2004) 347–365
6. Giunchiglia, F., Marchese, M., Zaihrayeu, I.: Encoding classifications into lightweight ontologies. JoDS **8** (2007) 57–81

7. Sabou, M., Gracia, J.: Spider: Bringing non-equivalence mappings to OAEI. In: Proc. of the Third International Workshop on Ontology Matching. (2008)
8. Noy, N.F., Musen, M.A.: Evaluating ontology-mapping tools: Requirements and experience. In: Proc. of OntoWeb-SIG3 Workshop. (2002) 1–14
9. Do, H.H., Melnik, S., Rahm, E.: Comparison of schema matching evaluations. In: Proc. of the 2nd International Workshop on Web Databases. (2002) 221–237
10. Kaza, S., Chen, H.: Evaluating ontology mapping techniques: An experiment in public safety information sharing. Decision Support Systems **45**(4) (2008) 714–728
11. Isaac, A., Wang, S., Zinn, C., Matthezing, H., van der Meij, L., Schlobach, S.: Evaluating thesaurus alignments for semantic interoperability in the library domain. IEEE Intelligent Systems **24**(2) (2009) 76–86
12. Ehrig, M., Euzenat, J.: Relaxed precision and recall for ontology matching. In: Proc. of Integrating Ontologies Workshop. (2005)
13. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: Proc. of IJCAI. (2007) 348–353
14. David, J., Euzenat, J.: On fixing semantic alignment evaluation measures. In: Proc. of the Third International Workshop on Ontology Matching. (2008)
15. Meilicke, C., Stuckenschmidt, H.: Incoherence as a basis for measuring the quality of ontology mappings. In: Proc. of the 3rd International Workshop on Ontology Matching. (2008)
16. van Hage, W.R., Isaac, A., Aleksovski, Z.: Sample evaluation of ontology-matching systems. In: Proc. of EON. (2007) 41–50
17. Giunchiglia, F., Yatskevich, M., Avesani, P., Shvaiko, P.: A large dataset for the evaluation of ontology matching systems. The Knowledge Engineering Review Journal **24** (2008) 137–157
18. Caracciolo, C., Stuckenschmidt, H., Svab, O., Svatek, V., Euzenat, J., Hollink, L., Ichise, R., Isaac, A., Malaisé, V., Meilicke, C., Pane, J., Shvaiko, P.: First results of the ontology alignment evaluation initiative 2008. In: Proc. of the Third International Workshop on Ontology Matching. (October 2008)
19. Avesani, P., Giunchiglia, F., Yatskevich, M.: A large scale taxonomy mapping evaluation. In: Proc. of ISWC. (2005) 67–81
20. Melnik, S., Garcia-Molina, H., Rahm, E.: Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In: Proc. of ICDE. (2002)
21. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic schema matching with cupid. In: VLDB. (2001) 49–58
22. Do, H.H., Rahm, E.: Coma — a system for flexible combination of schema matching approaches. In: VLDB. (2002) 610–621
23. Spiliopoulos, V., Valarakos, A.G., Vouros, G.A.: CSR: Discovering subsumption relations for the alignment of ontologies. In: Proc. of ESWC. (2008) 418–431
24. David, J., Guillet, F., Briand, H.: Association rule ontology matching approach. International Journal on Semantic Web and Information Systems **3**(2) (2007) 27–49
25. Bouquet, P., Serafini, L., Zanobini, S.: Semantic coordination: A new approach and an application. In: Proc. of ISWC. (2003) 130–145
26. Giunchiglia, F., Maltese, V., Autayeu, A.: Computing minimal mappings. Technical report, University of Trento, DISI (2008)
27. Giunchiglia, F., Soergel, D., Maltese, V., Bertacco, A.: Mapping large-scale knowledge organization systems. In: Proc. of ICSD. (2009)