Results of the Ontology Alignment Evaluation Initiative 2006 ⋆

Jérôme Euzenat[1], Malgorzata Mochol[2], Pavel Shvaiko[3], Heiner Stuckenschmidt[4], Ondřej Šváb[5], Vojtěch Svátek[5], Willem Robert van Hage[6], and Mikalai Yatskevich[3]

[1] INRIA Rhône-Alpes, Montbonnot, France
jerome.euzenat@inrialpes.fr
[2] Free University of Berlin, Berlin, Germany
mochol@inf.fu-berlin.de
[3] University of Trento, Povo, Trento, Italy
{pavel,yatskevi}@dit.unitn.it
[4] University of Mannheim, Mannheim, Germany
heiner@informatik.uni-mannheim.de
[5] University of Economics, Prague, Czech Republic
{svabo,svatek}@vse.cz
[6] Vrije Universiteit Amsterdam, The Netherlands
wrvhage@few.vu.nl

Abstract. We present the Ontology Alignment Evaluation Initiative 2006 campaign as well as its results. The OAEI campaign aims at comparing ontology matching systems on precisely defined test sets. OAEI-2006 built over previous campaigns by having 6 tracks followed by 10 participants. It shows clear improvements over previous results. The final and official results of the campaign are those published on the OAEI web site.

1 Introduction

The Ontology Alignment Evaluation Initiative[7] (OAEI) is a coordinated international initiative that organizes the evaluation of the increasing number of ontology matching systems. The main goal of the Ontology Alignment Evaluation Initiative is to be able to compare systems and algorithms on the same basis and to allow anyone for drawing conclusions about the best matching strategies. Our ambition is that from such evaluations, tool developers can learn and improve their systems. The OAEI campaign is the evaluation of matching systems on consensus test cases.

Two first events have been organized in 2004: (i) the Information Interpretation and Integration Conference (I3CON) held at the NIST Performance Metrics for Intelligent Systems (PerMIS) workshop and (ii) the Ontology Alignment Contest held at the Evaluation of Ontology-based Tools (EON) workshop of the annual International

⋆ This paper improves on the "First results" initially published in the on-site Ontology matching workshop proceedings. The only official results of the campaign, however, are on the OAEI web site.

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

Semantic Web Conference (ISWC) [6]. The first unique OAEI evaluation campaign has been presented at the workshop on Integrating Ontologies held in conjunction with the International Conference on Knowledge Capture (K-Cap) 2005 [1]. The campaign of 2006 is presented at the Ontology Matching (OM) workshop at ISWC, in Athens, Georgia, USA.

In reaction over last year's remarks, this year we have a variety of test cases that emphasize different aspects of the matching needs. From three test cases last year, we now have six very different test cases. Some of these tests introduce particular modalities of evaluation, such as a consensus building workshop and application-oriented evaluation.

This paper serves as an introduction to the evaluation campaign of 2006 and to the results provided in the following papers. The remainder of the paper is organized as follows. In Section 2 we present the overall testing methodology that has been used. Sections 3-8 discuss in turn the settings and the results of each of the test cases. Section 9 overviews lessons learned based on the campaign. Finally, Section 10 outlines future plans and Section 11 concludes.

2 General methodology

We present the general methodology for the 2006 campaign as it was defined and report its execution.

2.1 Test cases

This year's campaign has consisted of four tracks gathering six data sets and different evaluation modalities.

Comparison track: benchmark (§3) Like in previous campaigns, a systematic benchmark series has been produced. The goal of this benchmark series is to identify the areas in which each matching algorithm is strong or weak. The test is based on one particular ontology dedicated to the very narrow domain of bibliography and a number of alternative ontologies of the same domain for which alignments are provided.

Expressive ontologies

anatomy (§4) The anatomy real world case covers the domain of body anatomy and consists of two ontologies with an approximate size of several 10k classes and several dozens of relations.

jobs (§5) The jobs test case is an industry evaluated real world business case. A company has a need to improve job portal functionality with semantic technologies. To enable higher precision in retrieval of relevant job offers or applicant profiles, OWL ontologies from the employment sector are used to describe jobs and job seekers and matching with regard to these ontologies provides the improved results. For confidentiality reasons, the test is run by the company team with software provided by the participants.

Directories and thesauri

directory (§6) The directory real world case consists of matching web directories, such as open directory, Google and Yahoo. It has more than 4 thousands of elementary tests.

food (§7) Two SKOS thesauri about food have to be aligned using relations from the SKOS Mapping vocabulary. All results are evaluated by domain experts. Each participant is asked to evaluate a small part of the results of the other participants.

Consensus workshop: conference (§8) Participants have been asked to freely explore a collection of conference organization ontologies (the domain being well understandable for every researcher). This effort was expected to materialize in usual alignments as well as in interesting individual correspondences ("nuggets"), aggregated statistical observations and/or implicit design patterns. There is no a priori reference alignment. For a selected sample of correspondences, consensus was sought at the workshop and the process of reaching consensus was recorded.

Table 1 summarizes the variation in the results expected from these tests.

test	language	relations	confidence	modalities
benchmarks	OWL	=	[0 1]	open
anatomy	OWL	=	1	blind
jobs	OWL	=	[0 1]	external
directory	OWL	=	1	blind
food	SKOS	narrowMatch, exactMatch, broadMatch	1	blind+consensual
conference	OWL-DL	=, $\leq$	1	blind+consensual

Table 1. Characteristics of test cases (open evaluation is done with already published expected results, blind evaluation is done by organizers from reference alignments unknown to the participants, consensual evaluation is obtained by reaching consensus over the found results and external evaluation is preformed independently of the organizers by running the actual systems).

2.2 Preparatory phase

The ontologies and alignments of the evaluation have been provided in advance during the period between June 1st and June 28th. This gave potential participants the occasion to send observations, bug corrections, remarks and other test cases to the organizers. The goal of this preparatory period is to be sure that the delivered tests make sense to the participants. The tests still evolved after this period, but only for ensuring a better participation to the tests. The final test base has been be released on August 23rd.

2.3 Execution phase

During the execution phase the participants used their systems to automatically match the ontologies from the test cases. Participants have been asked to use one algorithm

and the same set of parameters for all tests in all tracks. It is fair to select the set of parameters that provide the best results (for the tests where results are known). Beside parameters, the input of the algorithms must be the two ontologies to be aligned and any general purpose resource available to everyone, i.e., no resource especially designed for the test. In particular, the participants should not use the data (ontologies and reference alignments) from other test sets to help their algorithms.

Ontologies are, in most cases, described in OWL-DL and serialized in the RDF/XML format. The expected alignments are provided in the Alignment format expressed in RDF/XML. All the participants also provided the papers that are published hereafter and a link to their program and its configuration parameters.

2.4 Evaluation phase

The organizers have evaluated the results of the algorithms used by the participants and provided comparisons on the basis of the provided alignments.

In order to ensure that it is possible to process automatically the provided results, the participants have been requested to provide (preliminary) results by September 4th. In the case of blind tests only the organizers did the evaluation with regard to the withheld reference alignments. In the case of double blind tests, the participants provide a version of their system and the values of the parameters if any.

The standard evaluation measures are precision and recall computed against the reference alignments. For the matter of aggregation of the measures we use weighted harmonic means (weights being the size of the true positives). This clearly helps in case of empty alignments. Another technique that has been used is the computation of precision/recall graphs so it was advised that participants provide their results with a weight to each correspondence they found.

New measures addressing some limitations of precision and recall have also been used for testing purposes. These were presented at the workshop discussion in order for the participants to provide feedback on the opportunity to use them in a further evaluation.

2.5 Comments on the execution

This year again, we had more participants than in previous years: 4 in 2004, 7 in 2005 and 10 in 2006. We also noted the increase in tools compliance and robustness: they had less problems to carry the tests and we had less problems to evaluate the results.

We have had not enough time so far to validate the results which have been provided by the participants. Last year, validating these results has proved feasible so we plan to do it again in the future (at least for those participants who provided their systems).

We summarize the list of participants in Table 2. Similar to last year not all participants provided results for all tests. They usually did those which are easier to run, such as benchmark, directory and conference. The jobs line corresponds to the participants who have provided an executable version of their systems. The variety of tests and the short time given to provide results have certainly prevented participants from considering more tests.

test \ system	falcon	hmatch	dssim	coma	automs	jhuapl	prior	RiMOM	OCM	nih	Total
benchmark	√	√	√	√	√	√	√	√	√		9
anatomy	√	√		√			√			√	5
jobs	√	√		√	√		√		√		6
directory	√	√		√	√		√	√	√		7
food	√	√		√			√	√			5
conference	√	√		√	√			√	√		6
certified											
confidence	√	√		√			√	√			5
time					√		√	√	√	√	5

Table 2. Participants and the state of their submissions. Confidence is ticked when given as non boolean value. Time indicates when participants included execution time with their tests.

Like last year, the time devoted for performing these tests (three months) and the period allocated for that (summer) is relatively short and does not really allow the participants to analyze their results and improve their algorithms. On the one hand, this prevents having algorithms to be particularly tuned for the tests. On the other hand, this can be frustrating for the participants. The timeline is very difficult to handle, hence, we should try to give more time for the next campaign.

The summary of the results track by track is provided in the following six sections.

3 Benchmark

The goal of the benchmark tests is to provide a stable and detailed picture of each algorithm. For that purpose, the algorithms are run on systematically generated test cases.

3.1 Test set

The domain of this first test is Bibliographic references. It is, of course, based on a subjective view of what must be a bibliographic ontology. There can be many different classifications of publications, for example, based on area and quality. The one chosen here is common among scholars and is based on publication categories; as many ontologies (tests #301-304), it is reminiscent to BibTeX.

The systematic benchmark test set is built around one reference ontology and many variations of it. The reference ontology is that of test #101.The participants have to match this reference ontology with the variations. These variations are focusing the characterization of the behavior of the tools rather than having them compete on real-life problems. The ontologies are described in OWL-DL and serialized in the RDF/XML format. This reference ontology contains 33 named classes, 24 object properties, 40 data properties, 56 named individuals and 20 anonymous individuals.

Since the goal of these tests is to offer some kind of permanent benchmarks to be used by many, the test is an extension of the 2004 EON Ontology Alignment Contest.

The reference ontology has been improved last year by including circular relations that were missing from the first test. In 2006, we have put the UTF-8 version of the tests as standard, the ISO-8859-1 being optional. Test numbering (almost) fully preserves the numbering of the first EON contest.

The kind of expected alignments is still limited: they only match named classes and properties, they mostly use the "=" relation with confidence of 1.

There are still three groups of tests in this benchmark:

- simple tests (1xx) such as comparing the reference ontology with itself, with another irrelevant ontology (the wine ontology used in the OWL primer) or the same ontology in its restriction to OWL-Lite;
- systematic tests (2xx) that were obtained by discarding some features from some reference ontology. It aims at evaluating how an algorithm behaves when this information is lacking. The considered features were:
  Name of entities that can be replaced by random strings, synonyms, name with different conventions, strings in another language than english,
  Comments that can be suppressed or translated in another language,
  Specialization Hierarchy ăthat can be suppressed, expanded or flattened,
  Instances that can be suppressed,
  Properties that can be suppressed or having the restrictions on classes discarded, and
  Classes that can be expanded, i.e., replaced by several classes or flattened.
- four real-life ontologies of bibliographic references (3xx) that were found on the web and left mostly untouched (there were added xmlns and xml:base attributes).

Full description of these tests can be found on the OAEI web site.

3.2 Results

Table 3 provides the consolidated results, by groups of tests. We display the results of participants as well as those given by some very simple edit distance algorithm on labels (edna). Like last year, the computed values are real precision and recall and not a simple average of precision and recall. The full results are on the OAEI web site.

These results show already that three systems are relatively close (coma, falcon and RiMOM). The RiMOM system is slightly ahead of the others on these raw results. The DSSim system obviously favoured precision over recall but its precision degrades with “real world” 3xx series. No system had strictly lower performance than edna.

The results have also been compared with the three measures proposed in [3] last year (symmetric, effort-based and oriented). These are generalisation of precision and recall in order to better discriminate systems that slightly miss the target from those which are grossly wrong. The three measures provide the same results. This is not really surprising given the proximity of these measures. As expected, they can only improve over traditional precision and recall. The improvement affects all the algorithms, but this is not always strong enough for being reflected in the aggregated results. Moreover, the new measures do not dramatically change the evaluation of the participating systems.

Each algorithm has its best score with the 1xx test series. There is no particular order between the two other series. Again, it is more interesting to look at the 2xx

algo	refalign		edna		automs		coma		DSSim		falcon		hmatch		jhuapl		OCM		prior		RiMOM	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	1.00	1.00	0.96	1.00	0.94	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.91	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00
2xx	1.00	1.00	0.90	0.49	0.94	0.64	0.96	0.82	0.99	0.49	0.91	0.85	0.83	0.51	0.20	0.86	0.93	0.51	0.95	0.58	0.97	0.87
3xx	1.00	1.00	0.94	0.61	0.91	0.70	0.84	0.69	0.90	0.78	0.89	0.78	0.78	0.57	0.18	0.50	0.89	0.51	0.85	0.80	0.83	0.82
Total	1.00	1.00	0.91	0.54	0.94	0.67	0.96	0.83	0.98	0.55	0.92	0.86	0.84	0.55	0.22	0.85	0.93	0.55	0.95	0.63	0.96	0.88
Symmetric	1.00	1.00	0.91	0.54	0.94	0.68	0.96	0.83	0.99	0.55	0.94	0.89	0.85	0.56	0.22	0.87	0.93	0.55	0.96	0.64	0.97	0.89
Effort-based	1.00	1.00	0.91	0.54	0.94	0.68	0.96	0.83	0.99	0.55	0.94	0.89	0.85	0.56	0.22	0.87	0.93	0.55	0.96	0.64	0.97	0.89
Oriented	1.00	1.00	0.91	0.54	0.94	0.68	0.96	0.83	0.99	0.55	0.94	0.89	0.85	0.56	0.22	0.87	0.93	0.55	0.96	0.64	0.97	0.89

Table 3. Means of results obtained by participants on the benchmark test case (corresponding to harmonic means).

series structure to distinguish the strengths of algorithms. This will be done in a separate document.

This year the apparently best algorithms provided their results with confidence measures. It is thus possible to draw precision/recall curves in order to compare them. We provide in Figure 1 the precision and recall graphs of this year. They involve only the results of participants who provided confidence measures different of 1 or 0 (see Table 2). They also feature the results for edit distances on class names (edna) and the results of Falcon last year (falcon-2005). The graph for falcon2005 is not really accurate since falcon2005 provided 1/0 alignments last year. This graph has been drawn with only technical adaptation of the technique used in TREC. Moreover, due to lack of time, these graphs have been computed by averaging the graphs of each of the tests (instead to pure precision and recall).

Contrary to last year, we have three systems competing at the higest level (falcon, coma and RiMOM) and a gap between these and the next systems. No system is significantly outperformed by standard edit distance (edna). The best systems are at the level of last year's best system (falcon).

Like last year we have compared the results of this year's systems with the previous years on the basis of the 2004 tests. (Table 4). The three best systems (falcon, coma and RiMOM) arrive at the level of last year's best system (falcon). However, no system outperforms it.

Unfortunately no representant of the group of systems that followed falcon last year is present this year.

Year	2004				2005		2006					
System	fujitsu		stanford		falcon		RiMOM		falcon		coma	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2xx	0.93	0.84	0.98	0.72	0.98	0.97	1.00	0.98	0.97	0.97	0.99	0.97
3xx	0.60	0.72	0.93	0.74	0.93	0.83	0.83	0.82	0.89	0.78	0.84	0.69
H-means	0.88	0.85	0.98	0.77	0.97	0.96	0.97	0.96	0.97	0.95	0.98	0.94

Table 4. Evolution of the best scores over the years (on the basis of 2004 tests).

4 Anatomy

The focus of the anatomy test case is to confront existing matching technology with real world ontologies. Our aim is to get a better impression of where we stand with respect to really hard challenges that normally require an enormous manual effort and in-depth knowledge of the domain.

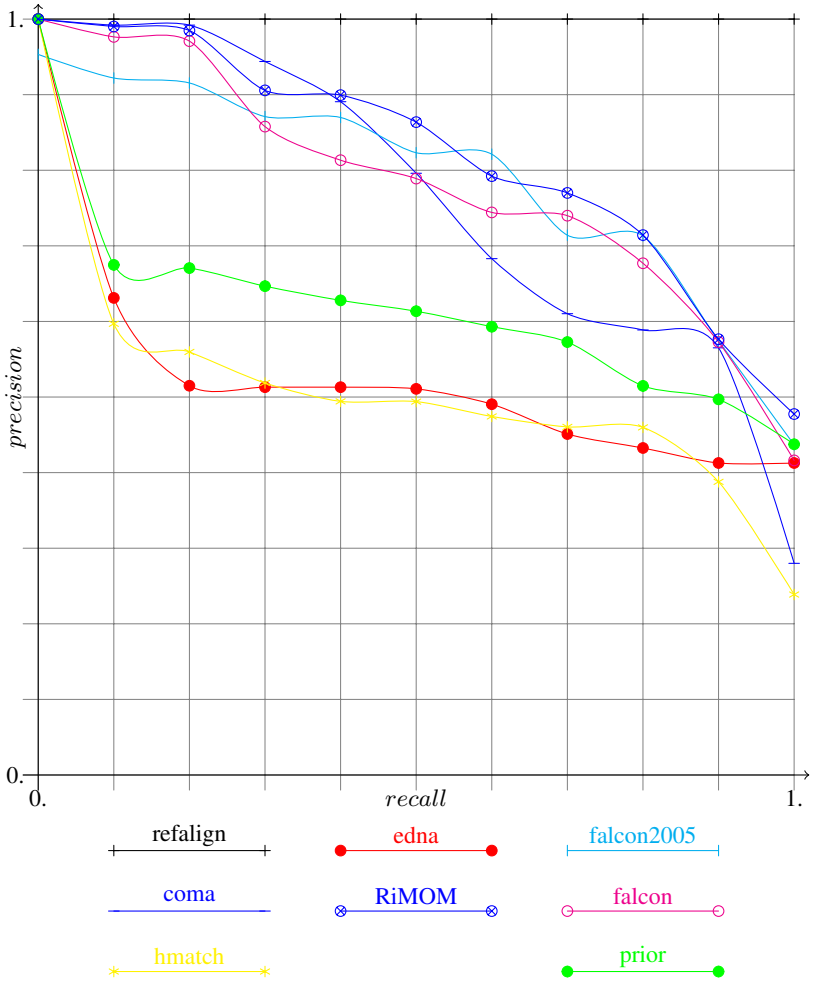

Fig. 1. Precision/recall graphs for the systems which provided confidence values in their results.

4.1 Test set

The task is placed in the medical domain as this is the domain where we find large, carefully designed ontologies. The specific characteristics of the ontologies are:

- Very large models: OWL models of more than 50MB.
- Extensive class hierarchies: Ten thousands of classes organized according to different views on the domain.
- Complex relationships: Classes are connected by a number of different relations.
- Stable terminology: The basic terminology is rather stable and should not differ too much in the different models.
- Clear modeling principles: The modeling principles are well-defined and documented in publications about the ontologies.

As a consequence, the anatomy test set actually tests existing matching systems with respect to two questions.

- Do existing approaches scale to very large models?
- Are existing approaches able to take advantage of well-documented modeling principles and knowledge about the domain?

This also means that the goal of this test case is not to compare the performance of matching systems on a quantitative basis. We are rather interested in how many systems are actually able to create mappings at all and in specific heuristics used for computing matches that are described in the corresponding papers.

The ontologies to be aligned are different representations of human anatomy developed independently by teams of medical experts. Both ontologies are available in OWL format and mostly contain classes and relations between them. The use of axioms is limited.

The Foundational Model of Anatomy The Foundational Model of Anatomy has been developed by the University of Washington. It is an ontology describing the human anatomy including a taxonomy of body parts, information about anatomical structures and structure transformations. According to the developers the Foundational Model of Anatomy ontology contains approximately 75.000 classes and over 120.000 terms; over 2.1 million relationship instances from 168 relationship types link the FMA's classes into a coherent symbolic model.

We extracted an OWL version of the ontology from a Protégé database. The resulting model is in OWL-full as relations are defined between classes rather than instances.

Galen The second ontology is the anatomy model developed in the OpenGalen Project by the University of Manchester. According to the creators, the ontology contains around 10.000 concepts covering a bit more than standard textbook anatomy in terms of body parts, anatomical structures and relations between different parts and structures.

The ontology is freely available as a Protégé Project file on the OpenGalen web page. We created an OWL version of the ontology using the export functionality of Protégé. The resulting ontology is in OWL-DL thus supporting logical reasoning about inconsistencies.

4.2 Results

The anatomy use case is part of the ontology alignment evaluation challenge for the second time now. While in 2005, none of the participants was in the position to submit a result for this data set. Almost all participants reported major difficulties in processing the ontologies due to their size and the fact that one of the models is in OWL full. At least these scalability problems seem to be solved this year. In the 2006 campaign, five out of ten participants submitted results for the anatomy data set. This clearly shows the advance of matching systems on the technical level and also shows that matching technologies are ready for large scale applications.

On the content level the results are much harder to judge. Due to the lack of a reference mapping, we were not able to provide a quantitative judgement and comparison of

the different systems. In the evaluation, we rather concentrated on the coverage of the ontologies, the degree of agreement amongst the matching systems and on the specific techniques used the matching systems to address this challenging alignment task.

A first observation, we made was that none of the systems managed to reach a good coverage of the ontologies. Although both models contain several ten thousand concepts and the fact that we can assume a high degree of overlap in the two models, the systems were only able to produce mappings for 2000 to 3000 concepts, which is less than 5% of the concepts in the FMA.

We also found out that systems have severe difficulties with irregular concept names. The GALEN ontology contains a subset of concepts with highly irregular concept names. It turned out that only one system (Coma++) was able to determine mappings for these concepts – for the price of not being able to match any of the concept names with regular names.

Looking at the actual methods used by the systems we see a common pattern. Almost all systems use the linguistic similarity between class names and other features of the class description as a basis for determining candidates. Normally, the systems combine different similarity measures. On top of this purely linguistic comparison, some systems also apply structural techniques. In particular, they translate the models into a graph structure and propagate the individual similarity in the graph structure. Only one of the systems (NIH) actually used reasoning techniques to validate hypothesis and to determine matches based on the semantics of the models.

4.3 Discussion and Conclusions

For the first time since the anatomy data set has been used in the ontology alignment evaluation challenge, we are in a position, were we were actually able to compare the results of different matching systems. The results show that there is still a lot of work to do to make matching systems ready for real life applications. The problems above showed that differences in the naming scheme of classes can already cause matchers to fail on a significant subset of the vocabulary. It seems that existing matchers suffer from the need to balance precision and recall in determining mappings. This conclusion is backed by results from other experiments, where it turned out that matching systems that produce highly precise mappings miss many mappings found by other systems. We conclude that using fixed thresholds to determine mapping candidates is not a good way for trading-off precision and recall.

We were disappointed to see that only one system actually used some form of reasoning in order to take the meaning of the ontologies into account. As one of the major advantages of OWL is the ability to specify and reason about the semantics of concepts, it is at least surprising that this feature is not exploited by existing matchers. In fact, logical reasoning could be a way of becoming less dependent on the quality of certain similarity measures that obviously have some limitations when it comes to complex ontologies.

In summary, the results of the anatomy test case have shown that there is some significant progress in terms of the maturity of matching technology. On the other hand, the results also show that there are still a lot of open problems with respect to producing good alignments on real life cases. For the setup of the next challenge this means that

we have to think about a more precise evaluation of the matching results in order to determine where exactly the problems of different matchers are. For this purpose it is necessary to have a reference alignment to compare against. Currently there are two possible ways to make such an evaluation possible. The first option is to move to a different, but related data set for which a reference mapping exists. Such a data set exists in terms of the anatomy part of the NCI thesaurus and the Adult Mouse Anatomy ontology. These ontologies would be much smaller in size but support a quantitative evaluation. The other option is to start building a reference mapping for the current data set using the mappings created by the participants of this years challenge as a starting point.

5 Jobs

The goal of the job test case is to evaluate the results of matching ontologies in the application context.

5.1 Test set

Semantic web technologies are used to semantically annotate job postings and applicant profiles in order to increase market transparency together with avoiding the bottleneck of a central database. In a semantic recruitment application the data exchange between employers, job applicants and job portals is based on a set of shared vocabularies describing domain relevant terms: occupations, industrial sectors and skills. These commonly used vocabularies have been formally defined by means of a Human Resource ontology (HR-ontology). The implementation of the HR-ontology was realized by translating several semi-structured input formalisms and encoding text-based classification standards into OWL. This ontology is used in a job matching application for computing the similarity between jobs and profiles. The current application uses an algorithm which is based on the similarity between two concepts determined by the distance between them.

We planned to modify this application in order that it can take advantage of the alignments found by participants to compute similarity. The matching systems as well as their parameters have been provided to the organizers who could run the algorithms on the ontologies and obtain the alignments. These alignments were to be used by job matchers in order to compare 250 job offers with about 250 applicant profiles.

5.2 Results

The results are not available at the time of writing. They will be made available on the OAEI web site if we can find time to complete this test.

6 Directory

The directory test case aims at providing a challenging task for ontology matchers in the domain of large directories.

6.1 Test set

The data set exploited in the web directories matching task was constructed from Google, Yahoo and Looksmart web directories as described in [2; 7]. The dataset is presented as taxonomies where the nodes of the web directories are modeled as classes and classification relation connecting the nodes was modeled as rdfs:subClassOf relation.

The key idea of the data set construction methodology is to significantly reduce the search space for human annotators. Instead of considering the full mapping task which is very big (Google and Yahoo directories have up to $3 * 10^5$ nodes each: this means that the human annotators need to consider up to $(3 * 10^5)^2 = 9 * 10^{10}$ mappings), it uses semi automatic pruning techniques in order to significantly reduce the search space. For example, for the dataset described in [2], human annotators consider only 2265 mappings instead of the full mapping problem.

So, there has been 3 proposed representation for this year:

- one matching task between two taxonomies of 10^3 categories (full test set),
- one matching task between two taxonomies of 10^2 categories (10% test set), and
- 4639 matching task between two paths of around 10 categories (unit test sets).

The first data set incorporates the matching tasks involved in the unit tests which also correspond to the reference set. The second data set is guarantee to contain 10% of these unit tests.

This year the reference data set has been significantly extended with respect to the one exploited in OAEI-2005 [4]. In particular, the reference mapping contains not only positive but also negative mappings, which are used to approximate not only recall but also precision. The key difference of the reference mapping with respect to conventional ones, such as ones exploited in benchmark tests in this evaluation, is that it does not contain the complete set of reference mappings (R). Instead of this the reference mapping is composed of two parts [7]:

- Representative subset of complete reference mapping ($P \subseteq R$). It contains the positive mappings, i.e., the mappings that hold for the matching task.
- Representative subset of negative mappings ($N \subseteq \bar{R}$), i.e., the mappings that do not hold for the matching task.

The reference mapping is composed of 2265 positive and 2374 negative mappings. Therefore the matching unit test set corresponds to 2265+2374=4639 tasks of finding the semantic relation holding between paths to root in the web directories modeled as sub class hierarchies.

6.2 Results

Approximate precision, recall and F-measure of the systems on web directories dataset are presented on Figure 2, 3 and 4 respectively. Given an alignment A and the set P and N of positive and negative mappings, approximate precision and recall are computed

by:

$$AP(A,P,N) = \frac{|P \cap A|}{|P \cap A| + |N \cap A|} \qquad AR(A,P) = \frac{|P \cap A|}{|P|}$$

These formula, especially that of AP, generalize precision and recall by not taking the whole set of valid correspondences as reference alignment. They are called approximate precision and recall because when $P = R$ and $N = \bar{R}$ (R is the reference alignment), they correspond to precision and recall. How this is an accurate approximation of precision and recall heavily depends on the choice of P and N; these are discussed in [2; 7].

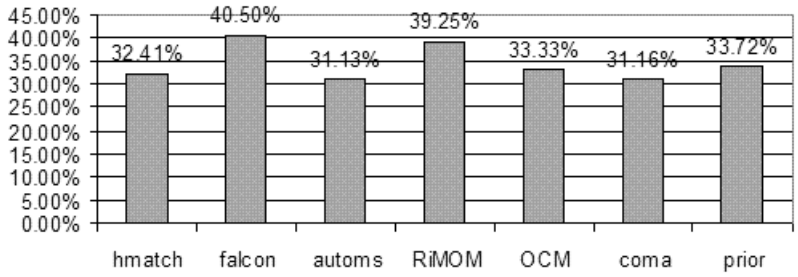

Fig. 2. Approximate precision for web directories matching task.

Similarly with OAEI-2005, 7 matching systems were evaluated on the dataset. However, only one of them (Falcon) participated in both evaluations. The systems in general demonstrated higher results than in OAEI-2005. The average approximate recall of the systems increased from 22.23% to 25.82%. The highest approximate recall (45.47%) was demonstrated by the Falcon system what is almost 50% increase in respect to its last year result (31.17%).

Despite of this progress the dataset remains difficult for the matching systems. The maximum and average values for approximate precision (40.5% and 34.5%), approximate recall (45.47% and 25.82%) and approximate F-measure (42.85% and 28,56%) are significantly lower than corresponding real values in benchmark tests for example.

Partition of positive and negative mappings according to the systems results are presented on Figure 5 and 6.

As from the figures, 43% of positive mappings have not been found by any of the systems. At the same time 22% of negative mappings were found by all the matching systems, i.e., all the matching systems mistakenly returned them as positive. Moreover only 10% of positive mappings were found by all the matching systems.

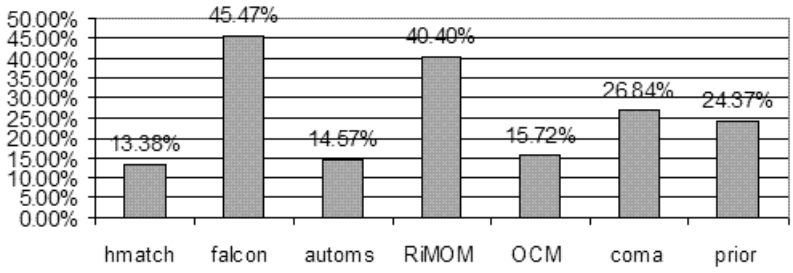

Fig. 3. Approximate recall for web directories matching task.

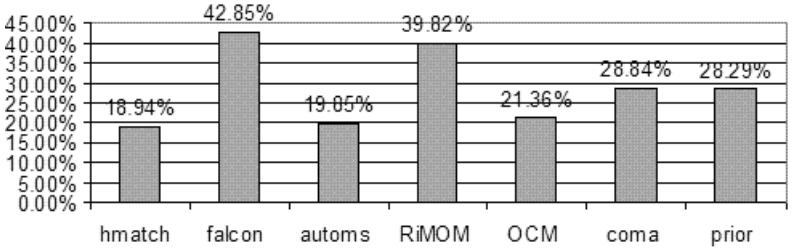

Fig. 4. Approximate F-measure for web directories matching task.

6.3 Comments

Six out of seven systems that participated in the evaluation presented their results only for one of the dataset representations, namely for the representation composed from 4639 node matching tasks. Only one system (H-Match) presented the results also for the other representations. Since, the other tasks were proposed in order to test scalability of the approaches, this can be interpreted as a sign of poor scalability of the systems participating in the evaluation.

Blind evaluation declined for some of the systems a possibility to improve their final results after preliminary result disclosure. For example, the final results of the coma and prior matching systems were slightly lower than their preliminary results. The final F-measure of coma dropped from 32.56% to 28.84% while F-measure of prior dropped from 28.32% to 28.29%.

7 Food

The food test case is another taxonomy task in which the taxonomies are taken out of theauri, i.e., they have a lot of text involved compared to the previous test case, and they are expressed in SKOS.

7.1 Test set

The task of this case consists of matching two thesauri formulated in SKOS:

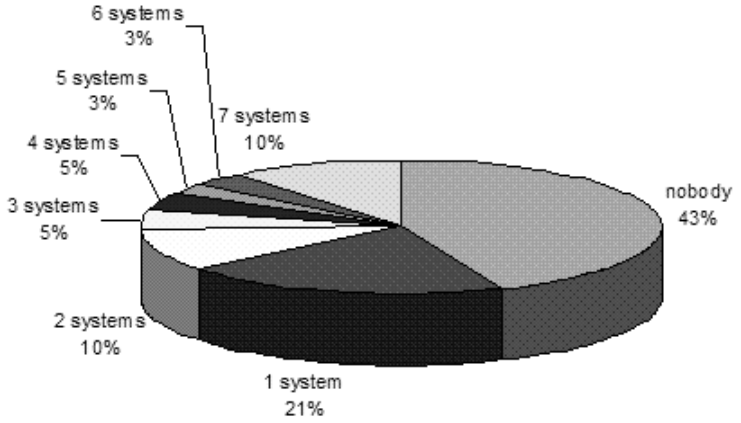

Fig. 5. Partition of the systems results on positive mappings.

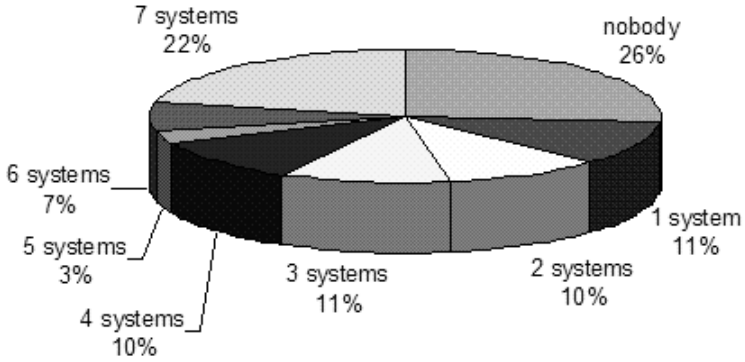

Fig. 6. Partition of the systems results on negative mappings.

AGROVOC The United Nations Food and Agriculture Organization (FAO) AGROVOC thesaurus, version May 2006. This thesaurus consists of 28.174 descriptor terms, i.e., prefered terms, and 10.028 non-descriptor terms, i.e., alternative terms. AGROVOC is multilingual in ten languages (en, fr, es, ar, zh, pt, cs, ja, th, sk).

NALT The United States National Agricultural Library (NAL) Agricultural thesaurus, version 2006. This thesaurus consists of 41.577 descriptor terms and 24.525 non-descriptor terms. NALT is monolingual, English.

Participants had to match these SKOS versions of AGROVOC and NAL using the exactMatch, narrowMatch, and broadMatch relations from the SKOS Mapping Vocabulary.

7.2 Evaluation procedure

Five participants took part in the OAEI 2006 food alignment task, South East University (Falcon-AO), University of Pittsburgh (Prior), Tsinghua University (RiMOM), Uni-

versity of Leipzig (COMA++), and Universitá degli Studi di Milano (HMatch). Each team provided between 10.000 and 20.000 alignments. This amounted to 31.112 unique alignments in total.

In order to give dependable precision results within the time span of the alignment initiative given a limited number of assessors we did a sample evaluation on 7% of the alignments. This sample was chosen to be representative of the type of topics covered by the thesauri and to be impartial to each participant and impartial to how much consensus amongst the participants there was about each alignment. We distinguished three categories of topics in the thesauri that each required a different level of domain knowledge of the assessors: Taxonomical concepts (plants, animals, bacteria, etc.), biological and chemical terms (structure formulas, terms from generics, etc.), and the remaining concepts (geography, agricultural processes, etc.). Under the authority of taxonomists at the US Department of Agriculture the taxonomical category of mappings was assessed using the strict rules that apply to the naming scheme of taxonomy. These are that if the preferred term of one concept is exactly the same as either the preferred or the alternative term of another concept then the concepts are considered to be exact matches. The latter two categories were assessed by two groups, a group of domain experts from the USDA and the FAO, and a group of computer scientists at the EKAW conference. The agreement between these groups was 72%. The computer scientists were less likely to judge an alignment to be correct than the domain experts.

As a significance test on precision scores of the systems we used the Bernoulli distribution. The precision of system A, P_A can be considered to be significantly greater than that of system B for a sample set of size N (in the cases of the three categories we distinguished respectively 18.399, 250, and 650) when the following formula holds:

$$|P_A - P_B| > 2\sqrt{\left(\frac{P_A \cdot (1 - P_A)}{N}\right)^2 + \left(\frac{P_B \cdot (1 - P_B)}{N}\right)^2}$$

Giving dependable recall numbers within the time span of the alignment initiative was not feasible, so we estimated recall on four sample sub-hierarchies of the thesauri: All oak trees (everything under the concept representing the Quercus genus), All rodents (everything under Rodentia), Geographical concepts of Europe, and everything under the NALT concept animal health and all AGROVOC concepts that have alignments to these concepts and their sub-concepts. These four samples respectively have size 41, 42, 74, and 34. Around 30% of the mappings were broadMatch and narrowMatch, the rest was exactMatch.

7.3 Results

The taxonomical parts of the thesauri accounted for by far the largest part of the alignments. The more difficult alignments that required lexical normalization, such as structure formulas, and relations that required background knowledge, such as many of the relations in the miscellaneous domain, accounted for a smaller part of the alignment. This caused systems that did well at the taxonomical mappings to have a great advantage over the other systems. The Falcon-AO system performed consistently best at the largest of the two categories and thus achieved high precision.

All systems only returned exactMatch alignments. This means that recall of all systems was limited to 71%. The RiMOM system managed to discover more good results than the Falcon-AO system on the four small sample recall bases, at the cost of some precision. Since recall was assessed on such a small set of examples we can only draw conclusions based on the precision results, but if the difference in recall between RiMOM and Falcon-AO persists throughout the rest of them, RiMOM achieves a better F-measure than Falcon-AO.

	RiMOM	Falcon-AO	Prior	COMA++	HMatch
Precision (taxonomical)	82%	83%⋆	68%	43%	48%
Precision (bio/chem)	85%⋆	80%	81%	76%	83%
Precision (miscellaneous)	78%	83%⋆	74%	70%	80%
Precision (all topics)	**81%**	**83%⋆**	**71%**	**54%**	**61%**

Table 5. Precision results based on sample evaluation. ⋆ indicates the significantly best system.

	RiMOM	Falcon-AO	Prior	COMA++	HMatch
Recall (all relations)	**50%**	**46%**	**45%**	**23%**	**46%**
Recall (only exactMatch)	71%	65%	64%	33%	65%

Table 6. Tentative estimation of recall based on sample evaluation.

	RiMOM	Falcon-AO	Prior	COMA++	HMatch
F-measure (all rel. & top.)	**62%**	**59%**	**55%**	**33%**	**53%**

Table 7. Tentative estimation of F-measure based on sample evaluation.

alignment found by # systems	1	2	3	4	5
average precision	6%	35%	67%	86%	99%
# alignments	21.663	2.592	2.470	4.467	5.555

Table 8. Consensus: average precision of the alignments returned by a number of systems.

8 Conference

The conference test set introduces matching several ontologies together as well as a consensus workshop aiming at studying the elaboration of consensus when establishing reference alignments.

8.1 Test set

The collection consists of ten ontologies in the domain of organizing conferences. The main features of this test set are:

- *Generally understandable domain.* Most ontology engineers are familiar with organizing conferences. Therefore, they can create their own ontologies as well as evaluate the mapping among their concepts with enough erudition.
- *Independence of ontologies.* Ontologies were developed independently and based on different resources, they thus capture the issues in organizing conferences from different points of view and with different terminology.
- *Relative richness in axioms.* Most ontologies were equipped with DL axioms of various kinds, which opens a way to use semantic matchers.

Ontologies differ in numbers of classes, of properties, in their DL expressivity, but also in underlying resources. Six ontologies are based *on tools* supporting the task of organizing conferences, two are based on experience of people with *personal participation* in conference organization, and two are based on *web pages* of concrete conferences.

8.2 Results

For the sake of brevity, all results from the initial evaluation phase are on the result report page[8]. There you can find global statistics about participants' results, which more or less reflect their quality. Additional, finer-grained results were obtained at the "consensus building workshop".

As there was no reference alignment to compare with, only a general statistics of submissions plus some simple observations were available to the date of writing this material. The statistics (counts of ontology pairs processed) follow:

- Automs tried to map all ontologies to three ontologies (30 alignments).
- Coma and Falcon delivered 45 alignments, i.e., all ontologies were mapped to each other.
- In the case of Hmatch, 90 pairs of ontologies were mapped, separately including each direction of mapping.
- RiMOM mapped all 100 pairs of ontologies, separately including each direction of mapping. Mapping of ontology onto itself was also included.
- In the case of OCM, 21 pairs of ontologies were mapped. Some ontologies were omitted because of their high complexity.

[8] http://nb.vse.cz/ svabo/oaei2006/

Other comments:

- Only equivalence, i.e., no subsumption, relations were discovered; for concepts and for properties separately, not across.
- Four participants delivered correspondences with certainty factors between 0 and 1 (coma, falcon, hmatch and RiMOM); the two remaining ones (automs and OCM) delivered 'certain' correspondences.
- One associated OAEI paper, by the AUTOMS team, discussed the role of different techniques (string matching, structure matching, thesaurus term matching) for different correspondences.
- Independently from OAEI, the conference collection has been investigated with the method of [5]. They evaluated the alignments of four systems, only among concepts. On the base of their evaluation, Falcon outperforms others in terms of precision (based on single person judgment). Their evaluation was also discussed at the consensus workshop.

Consensus building workshop During the "Ontology matching" workshop we organized a "Consensus building workshop". The main idea behind this workshop was to thoroughly discuss controversial mappings, i.e., those on which tools disagree, and thus partly provide *feedback* for authors of involved systems and partly examine the *argumentation process*. Altogether 9 mappings were discussed, chosen before the workshop by the organizers, and the group finally achieved consensus for each mapping. A presentation contains both the information about evaluation and discussed "controversial" mappings[9]. Chosen mappings as candidates of controversy were representatives of following phenomena:

- subsumption - Two elements are considered as equivalent by the systems, but they are rather in relation of subsumption, e.g., pairs: `Document` vs. `article` and `Location` vs. `Place`.
- inverse property - Pair of elements are considered as equivalent by systems, but they are inverse, e.g., `reviews` vs. `hasReview`.
- lexical confusion – This category contains such mappings considered by systems that are wrong and mainly based on lexical similarity, e.g., `PC_Member` vs. `Member_PC`.
- Other phenomena that were not included in the choice for discussion are for example siblings (elements are rather siblings) and heterogenous mappings (matching of relation to class or vice versa).

Regarding *arguments* against and for, we experience that *lexical* reasons of mapping are first considered. Then follow arguments with regard to *context* of elements in question. This means consideration of certain neighborhood, subclasses and superclasses (in the case of properties, we can consider subproperties and superproperties). This can disclose different extensions of classes (especially through their subclasses). Also, properties related to classes are considered. As a last resort, *axioms* (more complex restrictions) are taken into account if they are present.

[9] It can be downloaded from http://nb.vse.cz/ svabo/oaei2006/#organisation.

Discussion during the Consensus building workshop also showed us the necessity of considering mappings in the neighborhood and the possibility to build mappings from some mappings that are quite certain and they have 1:1 cardinality. Reaching consensus about mappings is not an easy process, but it is achievable if people can discuss the "controversial" issues based on the facts, i.e., ontology semantics and the full context of the mappings.

The process of analysing the results of participants also addressed critical remarks to our dataset. Some ontologies contain clear mistakes in terms of hierarchy of classes, naming elements (bad English) and do not fulfil some expectations made about this dataset like richness in axioms. On the other side, these features can be so widespread in ontologies available on the present and most certainly future semantic web that it makes this dataset a truly realcase.

9 Lesson learned

From last year's lesson learned, we have applied those concerning character encoding, new evaluation measures and having a progressive test suite in the directory case. However, we must admit that not all of them have been applied, partly due to lack of time. So we reiterate those lessons that still apply with new ones, including:

A) It is now a general trend that tools for the semantic web are more robust and compliant. As a consequence, we had comments on the tests this year that concerned problems not discovered in previous years. Obviously the tools can now better handle the ontologies proposed in the tests and they return results that are more easy to handle for the evaluation. Moreover, we had more participants able to handle large scale sets.
B) Not all the systems from the last year campaign participated in the campaign of this year. Fortunately, the best system participated this year as well. It will be useful to investigate if this is a definitive trend, whether we are evaluating research prototypes or "serious" systems.
C) The benchmark test case is not discriminant enough between systems. It is still useful for evaluating the strength and weakness of algorithms but does not seems to be sufficient anymore for comparing algorithms. We will have to look into better alternatives.
D) We have had more proposals for test cases this year (we had actively looked for them). However, the difficult lesson is that proposing a test case is not enough, there is a lot of remaining work in preparing the evaluation. Fortunately, with tool improvements, it will be easier to perform the evaluation.
E) It would be interesting and certainly more realistic, to provide some random gradual degradation of the benchmark tests (5% 10% 20% 40% 60% 100% random change) instead of a general discarding of a feature. This has not been done for reason of time.
F) Last but not least, as last year we must mention that the timeline for this evaluation is far from being ideal both from the participants and the evaluators points of view. More time must be allocated to this campaign next year.

10 Future plans

Future plans for the Ontology Alignment Evaluation Initiative are certainly to go ahead and to improve the functioning of the evaluation campaign. This involves:

- Finding new real world test cases;
- Improving the tests along the lesson learned;
- Accepting continuous submissions (through validation of the results);
- Improving the measures to go beyond precision and recall (we have done this for generalized precision and recall as well as for using precision/recall graphs, and will continue with other measures);
- Drawing lessons from the new test cases and establishing general rules for consensus reference and application-oriented evaluation.

Of course, these are only suggestions that will be refined during the coming year.

11 Conclusion

The tests that have been run this year were even more complete than those of the previous years. However, more teams participated and the results tend to be better. This shows that, as expected, the field of ontology matching is getting stronger (and we hope that evaluation has been contributing to this progress).

Reading the papers of the participants should help people involved in ontology matching to find what makes these algorithms work and what could be improved.

The Ontology Alignment Evaluation Initiative will continue these tests by improving both test cases and testing methodology for being more accurate. Further information can be found at:

`http://oaei.ontologymatching.org`.

Acknowledgments

We warmly thank each participant of this contest. We know that they have worked hard for having their results ready and they provided insightful papers presenting their experience. The best way to learn about the results remains to read what follows.

We are grateful to the teams of the University of Washington and the University of Manchester for allowing us to use their ontologies of anatomy.

We also thank the other members of the Ontology Alignment Evaluation Initiative Steering committee: Lewis Hart (AT&T, USA), Tadashi Hoshiai (Fujitsu, Japan), Todd Hughes (DARPA, USA), Yannis Kalfoglou (University of Southampton, UK), John Li (Teknowledge, USA), Natasha Noy (Stanford University, USA), York Sure (University of Karlsruhe, Germany) Raphaël Troncy (CWI, Amsterdam, The Netherlands), and Petko Valtchev (Université du Québec à Montréal, Canada).

This work has been partially supported by the Knowledge Web European network of excellence (IST-2004-507482). Ondřej Šváb and Vojtěch Svátek were also partially supported by IGA VSE grants no.26/05 “Methods and tools for ontological engineering” and no.12/06 “Integration of approaches to ontological engineering: design patterns, mapping and mining”.

References

1. Ben Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Proceedings of the K-Cap 2005 workshop on Integrating ontologies*, Banff (CA), 2005.
2. Paolo Avesani, Fausto Giunchiglia, and Mikalai Yatskevich. A large scale taxonomy mapping evaluation. In *Proc. 4th International Semantic Web Conference (ISWC)*, Galway (IE), 2005.
3. Marc Ehrig and Jérôme Euzenat. Relaxed precision and recall for ontology matching. In *Proc. K-Cap 2005 workshop on Integrating ontologies*, pages 25–32, Banff (CA), 2005.
4. Jérôme Euzenat, Heiner Stuckenschmidt, and Mikalai Yatskevich. Introduction to the ontology alignment evaluation 2005. In *Proc. K-Cap 2005 workshop on Integrating ontologies*, Banff (CA), 2005.
5. Christian Meilicke, Heiner Stuckenschmidt, and Andrei Tamilin. Improving automatically created mappings using logical reasoning. In *Proc. ISWC Ontology matching workshop*, Athens (GA US), 2006. this volume.
6. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the 3rd Evaluation of Ontology-based tools workshop (EON)*, Hiroshima (JP), 2004.
7. Mikalai Yatskevich, Fausto Giunchiglia, and Paolo Avesani. A large scale dataset for the evaluation of matching systems. Technical report, University of Trento, Povo (IT), 2005.

Grenoble, Berlin, Trento, Mannheim, Prague, and Amsterdam, December 3rd, 2006