First results of the Ontology Alignment Evaluation Initiative 2007 *

Jérôme Euzenat[1], Antoine Isaac[2], Christian Meilicke[3], Pavel Shvaiko[4], Heiner Stuckenschmidt[3], Ondřej Šváb[5], Vojtěch Svátek[5], Willem Robert van Hage[2], and Mikalai Yatskevich[4]

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

Abstract. We present the Ontology Alignment Evaluation Initiative 2007 campaign as well as its results. The OAEI campaign aims at comparing ontology matching systems on precisely defined test sets. OAEI-2007 builds over previous campaigns by having 4 tracks with 7 test sets followed by 17 participants. This is a major increase in the number of participants compared to the previous years. Also, the evaluation results demonstrate that more participants are at the forefront. The final and official results of the campaign are those published on the OAEI web site.

1 Introduction

The Ontology Alignment Evaluation Initiative[1] (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 provides the evaluation of matching systems on consensus test cases.

Two first events were organized in 2004: (i) the Information Interpretation and Integration Conference (I3CON) held at the NIST Performance Metrics for Intelligent

* This is only a preliminary version of the paper. It presents a partial and early view of the results. The final results will be published on the OAEI web site shortly after the ISWC+ASWC 2007 workshop on Ontology Matching (OM-2007) and will be the only official results of the campaign.

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

Systems (PerMIS) workshop and (ii) the Ontology Alignment Contest held at the Evaluation of Ontology-based Tools (EON) workshop of the annual International Semantic Web Conference (ISWC) [12]. Then, unique OAEI campaigns occurred in 2005 at the workshop on Integrating Ontologies held in conjunction with the International Conference on Knowledge Capture (K-Cap) [2] and in 2006 at the first Ontology Matching (OM) workshop collocated with ISWC [7]. Finally, in 2007, OAEI results are presented at the second Ontology Matching workshop collocated with ISWC+ASWC, in Busan, South Korea.

We have continued last year's trend by having a large variety of test cases that emphasize different aspects of the matching needs. We have kept particular modalities of evaluation for some of these test cases, such as a consensus building workshop.

This paper serves as an introduction to the evaluation campaign of 2007 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 3-9 discuss in turn the settings and the results of each of the test cases. Section 10 overviews lessons learned based on the campaign. Finally, Section 11 outlines future plans and Section 12 concludes.

2 General methodology

We present the general methodology for the 2007 campaign as it was defined and report its execution.

2.1 Tracks and test cases

This year's campaign has consisted of four tracks gathering seven data sets (one more than in 2006) and different evaluation modalities.

The benchmark track (§3): Like in previous campaigns, systematic benchmark series have 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.

The expressive ontologies track. Anatomy (§4): The anatomy real world case deals with matching the Adult Mouse Anatomy (2.744 classes) and the NCI Thesaurus (3.304 classes) describing the human anatomy.

The directories and thesauri track:

- **Directory (§5):** The directory real world case consists of matching web site directories (like open directory or Yahoo's). It has more than four thousands of elementary tests.
- **Food (§6):** Two SKOS thesauri about food have to be matched using relations from the SKOS Mapping vocabulary. Samples of the results are evaluated by domain experts.
- **Environment (§7):** Three SKOS thesauri about the environment have to be matched (A-B, B-C, C-A) using relations from the SKOS Mapping vocabulary. Samples of the results are evaluated by domain experts.

Library (§8): Two SKOS thesauri about books have to be matched using relations from the SKOS Mapping vocabulary. Samples of the results are evaluated by domain experts.

The conference track and consensus workshop (§9): 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. Organizers of this track offer a posteriori evaluation of results in part manually and in part by data-mining techniques. For a selected sample of correspondences, consensus will be sought at the workshop and the process of its reaching will be recorded.

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

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

Table 1. Characteristics of test cases (open evaluation is made with already published expected results, blind evaluation is made by organizers from reference alignments unknown to the participants, consensual evaluation is obtained by reaching consensus over the found results).

2.2 Preparatory phase

The ontologies and (where applicable) the alignments of the evaluation have been provided in advance during the period between May 15th and June 15th, 2007. 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 ensure that the delivered tests make sense to the participants. The final test base was released on July 2nd. The tests did not evolve after this period.

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 matched 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.

In most cases ontologies are described in OWL-DL and serialized in the RDF/XML format. The expected alignments are provided in the Alignment format expressed in RDF/XML [6]. Participants also provided the papers that are published hereafter and a link to their systems and their 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 3rd. In the case of blind tests only the organizers did the evaluation with regard to the withheld reference alignments.

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 the 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 further evaluations.

2.5 Comments on the execution

This year again, we had more participants than in previous years: 4 in 2004, 7 in 2005, 10 in 2006, and 17 in 2007. We can also observe a common trend: participants who keep on the developments of their systems improve the evaluation results over time.

We have had not enough time so far to validate the results which had been provided by the participants, but we scrutinized some of the results leading to improvements for some participants and retraction from other. Validating these results has proved feasible in the previous years so we plan to do it again in future.

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 variety of tests and the short time given to provide results have certainly prevented participants from considering more tests.

There are two groups of systems: those which can deal with large taxonomies (food, environment, library) and those which cannot. The two new test cases (environment and library) are those with the least number of participants. This can be explained by the size of ontologies or their novelty - there are no past results to compare with.

Software	confidence	*benchmark*	*anatomy*	*directory*	*food*	*environment*	*library*	*conference*
AgreementMaker	√		√					
AOAS	√		√					
ASMOV	√	√	√	√				√
DSSim	√	√	√	√	√	√	√	
Falcon-AO v0.7	√	√	√	√	√	√	√	√
Lily		√	√	√				√
OLA2	√	√		√				√
OntoDNA		√		√				√
OWL-CM		√						
Prior+	√	√	√	√	√			
RiMOM	√	√	√	√	√			
SAMBO		√	√					
SCARLET					√			
SEMA		√						√
Silas							√	
TaxoMap	√	√	√					
X-SOM	√	√	√	√	√			
Total=17	10	13	11	9	6	2	3	6

Table 2. Participants and the state of their submissions. Confidence stands for the type of result returned by a system: it is ticked when the confidence has been measured as non boolean value.

This year we have been able to devote more time to performing these tests and evaluation (three full months). This is certainly still too little especially during the summer period allocated for that. However, it seems that we have avoided the rush of previous years. 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 on 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, whose numbering it (almost) fully preserves. This year, no modification has been made since the last year benchmark suite.

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 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 features from some reference ontology. It aims at evaluating how an algorithm behaves when a particular type of 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* that can be suppressed, expanded or flattened;
- *Instances* that can be suppressed;

– *Properties* that can be suppressed or having the restrictions on classes discarded;
– *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).

After evaluation we have noted two mistakes in our test generation software, so that tests #249 and 253 still have instances in them. This problem already existed in 2005 and 2006. So the yearly comparison still holds. Full description of these tests can be found on the OAEI web site.

3.2 Results

13 systems participated in the benchmark track of this year's campaign. Table 3 provides the consolidated results, by groups of tests. We display the results of participants as well as those given by some simple edit distance algorithm on labels (edna). The computed values are real precision and recall and not an average of precision and recall. The full results are on the OAEI web site.

algo	edna		ASMOV		DSSim		Falcon		Lily		OLA2		OntoDNA	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	1.00
2xx	0.40	0.55	0.95	0.90	0.99	0.60	0.92	0.85	0.97	0.89	0.91	0.86	0.80	0.43
3xx	0.46	0.79	0.85	0.82	0.89	0.67	0.89	0.79	0.81	0.80	0.63	0.76	0.90	0.71
Total	0.44	0.60	0.95	0.90	0.98	0.64	0.92	0.86	0.96	0.89	0.89	0.87	0.83	0.49
Ext	0.59	0.80	0.97	0.92	0.99	0.64	0.96	0.89	0.97	0.90	0.93	0.90	Error	

algo	OWL-CM		Prior+		RiMOM		SAMBO		SEMA		TaxoMap		X-SOM	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.34	0.99	0.99
2xx	0.82	0.51	0.92	0.79	0.97	0.86	0.98	0.51	0.92	0.72	0.91	0.19	0.73	0.67
3xx	0.95	0.37	0.87	0.83	0.69	0.80	0.94	0.67	0.67	0.79	0.92	0.26	0.94	0.68
Total	0.85	0.54	0.93	0.81	0.95	0.87	0.98	0.56	0.90	0.74	0.92	0.21	0.76	0.70
Ext	Error		0.96	0.84	0.96	0.87	Error		0.93	0.77	Error		Error	

Table 3. Means of results obtained by participants on the benchmark test case (corresponding to harmonic means). The Ext line corresponds to the three extended precision and recall measures (see [5] and further explanations next).

These results show already that three systems are relatively ahead (ASMOV, Lily and RiMOM) with three close followers (Falcon, Prior+ and OLA2). No system had strictly lower performance than edna.

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 series structure to distinguish the strengths of algorithms.

The results have also been compared with the three measures proposed in [5] (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, so they have been displayed only once in Table 3 under the label "Ext". This is not really surprising given the proximity of these measures. As expected, they only improve over traditional precision and recall. Again, the new measures do not dramatically change the evaluation of the participating systems (all score are improved and the six leading systems are closer to each others). This indicates that the not immediately best systems (Falcon, OLA2) could certainly easily be corrected to reach the level of the best ones (RiMOM in particular). Since last year the implementation of the precision and recall evaluator has changed. As a consequence, a number of results which would have been rejected last year, and then corrected by the participants, were accepted this year. As a consequence, now, the extended precision and recall reject them: this concerns the systems marked with "Error".

This year the apparently best algorithms provided their results with confidence measures. It is thus possible to draw precision/recall graphs in order to compare them. We provide in Figure 1 the precision and recall graphs of this year. They are only relevant for the results of participants who provided confidence measures different of 1 or 0 (see Table 2). They also feature the results for edit distance on class names (edna) and the results of previous years (Falcon-2005 and RiMOM-2006). 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).

These results and those displayed in Figure 2 single out a group of systems, ASMOV, Lily, Falcon 0.7, OLA2, Prior+ and RiMOM which seem to perform these tests at the highest level of quality. Of these, ASMOV, Lily and RiMOM seem to have slightly better results than the three others.

Like the two previous years there is a gap between these systems and their followers. The good news is that one system (OLA2) has achieved to fill this gap without significantly changing its strategy[2].

We have compared the results of this year's systems with the results of the previous years on the basis of 2004 tests, see Table 4. The results of three best systems (ASMOV, Lily and RiMOM) are comparable but never identical to the results provided in the previous years by RiMOM (2006) and Falcon (2005). Like Falcon last year, RiMOM provided this year lower results than last year. Figure 1 shows that RiMOM has increased in precision and decreased in overall performance. There seems to be a limit that systems are not able to overcome. At the moment, it seems that these systems are at a level at which making more progress is very hard: we now have strong arguments that having a 100% recall and precision on all these tests is not a reachable goal.

[2] Disclosure: the author of these lines is a member of the OLA2 team.

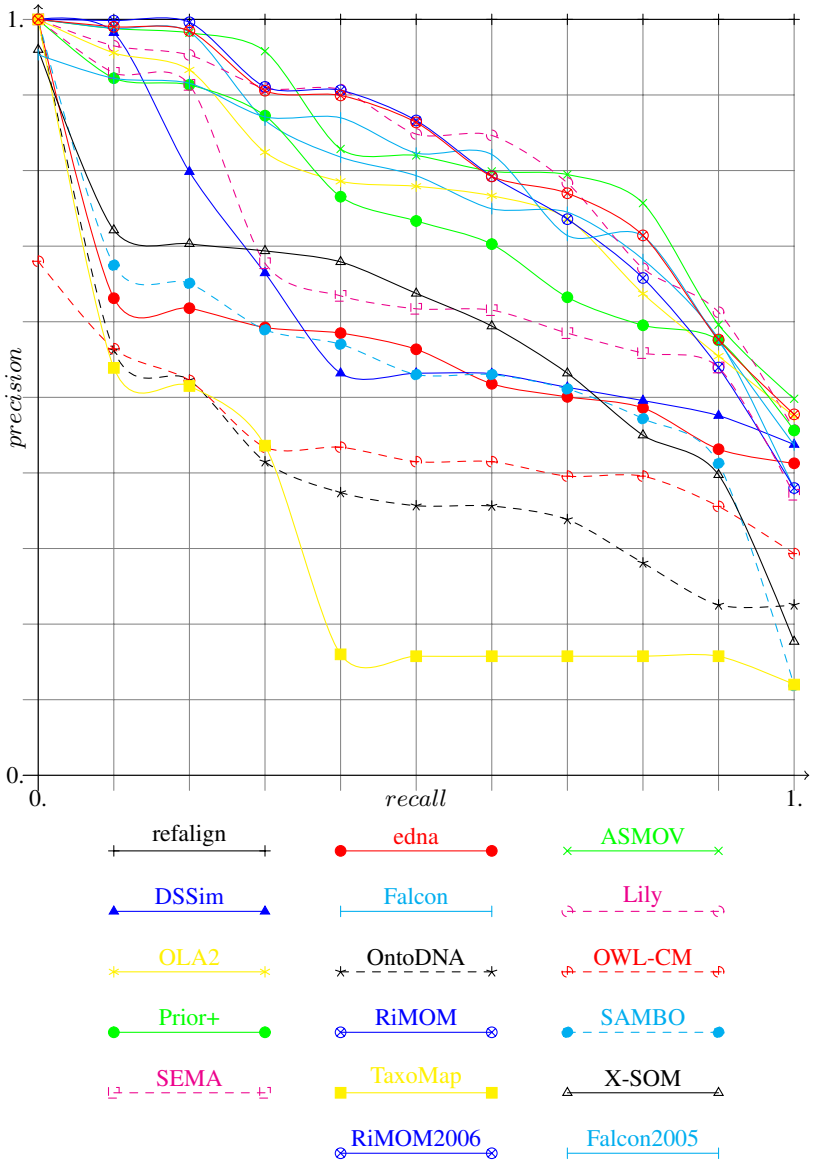

Fig. 1. Precision/recall graphs. They cut the results given by the participants under a threshold necessary for achieving $n\%$ recall and compute the corresponding precision. Systems for which these graphs are not meaningful (because they did not provide graded confidence values) are drawn in dashed lines. We remind the graphs for the best systems of the previous years, namely of Falcon in 2005 and RiMOM in 2006.

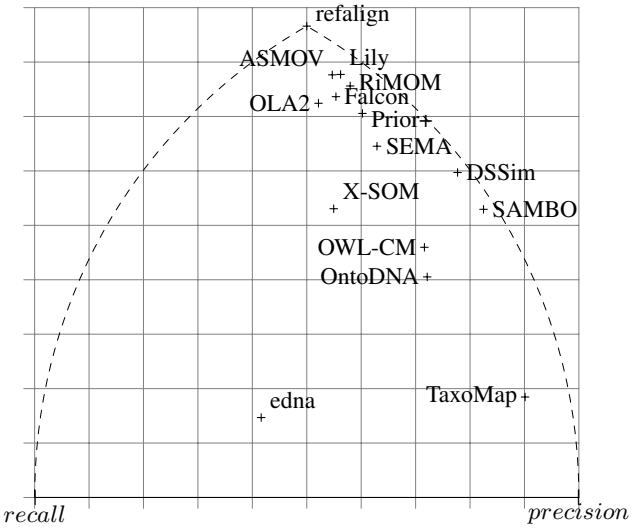

Fig. 2. Each point expresses the position of a system with regard to precision and recall.

Year	2004				2005		2006		2007					
System	Fujitsu		PromptDiff		Falcon		RiMOM		ASMOV		Lily		RiMOM	
test	Prec.	Rec.	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	1.00	1.00
2xx	0.93	0.84	0.98	0.72	0.98	0.97	1.00	0.98	0.99	0.99	1.00	0.98	1.00	0.97
3xx	0.60	0.72	0.93	0.74	0.93	0.83	0.83	0.82	0.85	0.82	0.81	0.80	0.69	0.80
H-means	0.88	0.85	0.98	0.77	0.97	0.96	0.97	0.96	0.97	0.97	0.97	0.96	0.95	0.95

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

4 Anatomy

The focus of the anatomy track is to confront existing matching technology with real world ontologies. Currently, we find such real world cases primarily in the biomedical domain, where a significant number of ontologies have been built covering different aspects of medical research. Manually generating alignments between these ontologies requires an enormous effort by highly specialized domain experts. Supporting these experts by automatically providing correspondence proposals is both challenging, due to the complexity and the specialized vocabulary of the domain, and relevant, due to the increasing number of ontologies used in clinical research.

4.1 Test data and experimental setting

The ontologies of the anatomy track are the NCI Thesaurus describing the human anatomy, published by the National Cancer Institute (NCI)[3], and the Adult Mouse Anatomical Dictionary[4], which has been developed as part of the Mouse Gene Expression Database project. Both resources are part of the Open Biomedical Ontologies (OBO). The complex and laborious task of generating the reference alignment has been conducted by a combination of computational methods and extensive manual evaluation. In addition, the ontologies were extended and harmonized to increase the number of correspondences between both ontologies. An elaborate description of creating the reference alignment can be found in [4] and in work to be published by Hayamizu et al.

The task is placed in a domain where we find large, carefully designed ontologies that are described in technical terms. Besides their large size and a conceptualization that is only to a limited degree based on the use of natural language, they also differ from other ontologies with respect to the use of specific annotations and roles, e.g., the extensive use of the $partOf$ relation. The manual harmonization of the ontologies leads to a situation, where we have a high number of rather trivial correspondences that can be found by simple string comparison techniques. At the same time, we have a good share of non-trivial correspondences that require a careful analysis and sometimes also medical background knowledge. To better understand the occurrence of non-trivial correspondences in alignments, we implemented a straightforward matching tool that compares normalized concept labels. This trivial matcher generates for all pairs of concepts $\langle C, D \rangle$ a correspondence if and only if the normalized label of C is identical to the normalized label of D. In general we expect an alignment generated by this approach to be highly precise while recall will be relatively low. With respect to our matching task we measured approximately 99% precision and 60% recall. Notice that the value for recall is relatively high, which is partially caused by the harmonization process mentioned above.

Because we assumed that all matchers would easily find the trivial correspondences, we introduce an additional measure for recall, called $recall+$. $Recall+$ measures how many non-trivial correct correspondences can be found in an alignment M. Given reference alignment R and alignment S generated by the naive string equality matching, $recall+$ is defined as follows:

[3] `http://www.cancer.gov/cancerinfo/terminologyresources/`

[4] `http://www.informatics.jax.org/searches/AMA_form..shtml`

$$Recall+ = \frac{|(R \cap M) - S|}{|R - S|}$$

We divided the task of automatically generating an alignment between these ontologies into three subtasks. Task #1 was obligatory for participants of the anatomy track, while task #2 and #3 were optional. For task #1 a matching system has to be applied with standard settings to obtain a result that is as good as possible with respect to the expected F-measure. For task #2 an alignment with increased precision has to be found. This seems to be an adequate requirement in a scenario where the automatically generated alignment will be directly used without subsequent manual evaluation. Contrary to this approach, in task #3 an alignment with increased recall has to be generated. Such an alignment could be seen as basis for subsequent expert evaluation. We believe that systems configurable with respect to these requirements will be much more useful in concrete application scenarios.

4.2 Results

In total, 11 systems participated in the anatomy task. These systems can be roughly divided in three groups. Systems of type A are highly specialized on matching biomedical ontologies and make extensive use of medical background knowledge. These systems are AOAS and SAMBO. Systems of type B can solve matching problems of different domains, but include a component exploiting biomedical background knowledge (e.g., using UMLS as lexical reference system). ASMOV and RiMOM fall into this category. Systems of type C, finally can be seen as general matching systems that do not distinguish between medical ontologies and ontologies of different domains. Most systems in the experiment fell into this category. Table 5 gives an overview of participating systems.

Runtime. The runtime of the systems differs significantly[5]. In average type-C systems outperformed systems that use medical knowledge. Falcon-AO, a system that solves large matching problems by applying a partition-based block matching strategy, solves the matching task in about 12 minutes without loss of quality with respect to the resulting alignment compared to other systems of type C. It has to be considered if similar approaches can also be applied to systems like SAMBO or Lily to solve their problems with runtime.

Type-C systems. The most astounding result is based on the suprisingly good performance of the naive label comparison approach compared to the alignments generated by systems of type C. The results of the naive approach are better with respect to recall as well as precision for testcase #1 compared to almost all matching systems of type C. Only TaxoMap and AgreementMaker generate alignments with higher recall but a significant loss in precision. We would have expected the participating systems

[5] Runtime information has been provided by the participants. All alignments have been generated on similarly equipped standard PCs. Advantages based on hardware differences could be neglected due to the significant differences in runtime.

System	Type	Testcase #1				Testcase #2		Testcase #3		Recall+	
		Runtime	Prec	Rec	F-meas	Prec	Rec	Prec	Rec	#1	#3
AOAS	A	n.a.	0.928	0.804	0.861	-	-	-	-	0.505	-
SAMBO	A	384 min	0.845	0.786	0.815	-	-	-	-	0.580	-
ASMOV	B	6 h	0.803	0.701	0.749	0.870	0.696	0.739	0.705	0.270	0.284
RiMOM	B	4 h	0.377	0.659	0.480	-	-	-	-	0.390	-
- Label Eq. -	-	*3 min*	*0.987*	*0.605*	*0.750*	-	-	-	-	0.0	-
Falcon-AO	C	12 min	0.964	0.591	0.733	0.986	0.540	0.814	0.655	0.123	0.280
TaxoMap	C	5 h	0.596	0.732	0.657	0.985	0.642	-	-	0.230	-
AgreementM.	C	30 min	0.558	0.635	0.594	0.930	0.286	0.424	0.651	0.262	0.302
Prior+	C	23 min	0.594	0.590	0.592	0.663	0.497	0.371	0.657	0.338	0.426
Lily	C	4 days	0.481	0.559	0.517	0.672	0.380	0.401	0.588	0.374	0.410
X-SOM	C	10 h	0.916	0.248	0.390	0.942	0.104	0.783	0.565	0.008	0.079
DSSim	C	75 min	0.208	0.187	0.197	-	-	-	-	0.067	-

Table 5. Participants and results with respect to runtime, precision, recall and F-measure. Results are listed in descending order with respect to the type of the system and the F-measure of testcase #1. The values for recall+ are presented in the rightmost columns for testcase #1 and #3.

to find more correct correspondences than applying straightforward label comparisons. It seems that many matching systems do not accept a correspondence even if the normalized labels of the concepts are equal. On the one hand, this might be caused by not detecting this equality at all (e.g., due to a partition based approach). On the other hand, a detected label equality can be rejected as correspondence due to the fact that additional information related to the concepts suggests that these concepts have a different meaning.

Type-A/B systems. Systems that use additional background knowledge related to the biomedical domain clearly generate better alignments compared to type-C systems. This result conforms with our expectations. The only exception is the low precision of the RiMOM system. The values for $recall+$ points to the advantage of using domain related background knowledge. Both AOAS and SAMBO detect about 50% of the non-trivial correspondences, while only Lily and Prior+ (systems of type C) achieve about 42% for testcase #3 with a significant loss in precision. Amongst all systems the AOAS approach generates the best alignment closely followed by SAMBO. Notice that AOAS is not available as a standalone system, but consists of a set of coupled programs which eventually require user configuration.

4.3 Discussion and conclusions

Obviously, the use of domain related background knowledge is a crucial point in matching biomedical ontologies and the additional effort of exploiting this knowledge pays off. This observation supports the claims for the benefits of using background knowledge made by other researchers [8; 1; 10]. Amongst all systems AOAS and SAMBO generate the best alignments, especially the relatively high number of detected non-trivial correspondences has to be mentioned positively. Nevertheless, for type C systems it is possible to detect non-trivial correspondences, too. In particular, the results of Lily and Prior+ on sub track #3 demonstrate this. Thus, there also seems to be a significant potential of exploiting knowledge encoded in the ontologies. Even if no medical background knowledge is used, it seems to make sense to provide a configuration that is specific to this type of domain. This is clearly demonstrated by the fact that most of the general matching systems fail to find a significant number of trivial correspondences. While in general it makes sense for a matcher not to accept all trivial correspondences to avoid the problem of homonymy, there are domains like the present one, however, where homonymy is not a problem, for example, because the terminology has been widely harmonized.

One major problem of matching medical ontologies is related to their large size. Though type C systems achieve relatively low values for recall, matching large ontologies seems to be less problematic. On the other hand the extensive use of domain related background knowledge has positive effects on recall, but does not seem to scale well. Thus, a trade-off between runtime and recall has to be found.

In further research we have to distinguish between different types of non-trivial correspondences. While for detecting some of these correspondences domain specific knowledge seems to be indispensable, the results indicate that there is also a large subset that can be detected by the use of alternative methods that solely rely on knowledge encoded in the ontologies. The distinction between different classes of non-trivial correspondences will be an important step for combining the strengths of both domain specific and domain independent matching systems. In summary, we can conclude that the data set used in the anatomy track is well suited to measure the characteristics of different matching systems with respect to the problem of matching biomedical ontologies.

5 Directory

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

5.1 Test set

The data set exploited in the directory matching task was constructed from Google, Yahoo and Looksmart web directories following the methodology described in [3; 9]. The data set is presented as taxonomies where the nodes of the web directories are modeled as classes and classification relation connecting the nodes is 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 matching task which is very large (Google and Yahoo directories have up to $3 * 10^5$ nodes each: this means that the human annotators need to consider up to $(3*10^5)^2 = 9*10^{10}$ correspondences), it uses semi automatic pruning techniques in order to significantly reduce the search space. For example, for the data set described in [3], human annotators consider only 2265 correspondences instead of the full matching problem.

The specific characteristics of the data set are:

- More than 4.500 of node matching tasks, where each node matching task is composed from the paths to root of the nodes in the web directories.
- Reference correspondences for all the matching tasks.
- Simple relationships, in particular, web directories contain only one type of relationships, which is the so-called classification relation.
- Vague terminology and modeling principles, thus, the matching tasks incorporate the typical real world modeling and terminological errors.

5.2 Results

The results are not available at the time of writing. They will be made available on the OAEI web site as soon as the evaluation is completed.

6 Food

The food test case is another taxonomy task in which the hierarchies come from theauri, i.e., they have a lot of text involved compared to the previous test case, and they are expressed in SKOS. Success in this task greatly depends on linguistic term disambiguation and recognition of naming conventions.

6.1 Test set

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

AGROVOC: The United Nations Food and Agriculture Organization (FAO) AGROVOC thesaurus, version February 2007. This thesaurus consists of 28.445 descriptor terms, i.e., prefered terms, and 12.531 non-descriptor terms, i.e., alternative terms. AGROVOC is multilingual in eleven languages (en, fr, de, es, ar, zh, pt, cs, ja, th, sk).

NALT: The United States National Agricultural Library (NAL) Agricultural thesaurus, version 2007. This thesaurus consists of 42.326 descriptor terms and 25.985 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.

6.2 Evaluation procedure

5 participants took part in the OAEI 2007 food matching task, South East University (Falcon-AO 0.7), University of Pittsburgh (PRIOR+), Tsinghua University (RiMOM), Politecnico di Milano (X-SOM), and the Knowledge Media Institute with two systems (DSSim and SCARLET). Each team provided between 18.420 (RiMOM) and 6583 (X-SOM) correspondences. This amounted to 37.384 unique correspondences in total.

In order to give dependable precision results within the time span of the OAEI campaign given a limited number of assessors we did a sample evaluation on roughly 4% 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 four 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.), geographical terms (countries, regions, etc.), and the remaining concepts (agricultural processes, natural resources, etc.). Under the authority of taxonomists at the US Department of Agriculture the taxonomical category of correspondences 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. This is possible due to common origins of the taxonomical parts of the thesauri. The alignments are currently being assessed by five groups of domain experts from the following institutions and companies: USDA NAL, UN FAO, Wageningen Agricultural University (WUR), Unilever, and the Netherlands organisation for applied scientific research (TNO).

As a significance test on precision scores of the systems we will use 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 when the following formula holds:

$$|P_A - P_B| > 2\sqrt{\frac{P_A(1-P_A)}{n} + \frac{P_B(1-P_B)}{n}}$$

Giving dependable recall numbers within the time span of the OAEI campaign will not be feasible, so we will estimate recall on a set of sample sub-hierarchies of the thesauri: Animal husbandry, fishery, 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.

6.3 Results

The results are not available at the time of writing. They will be made available on the OAEI web site as soon as the evaluation is completed.

7 Environment

The environment task is comprised of three matching tasks between three thesauri: the two thesauri of the food task (AGROVOC and NALT), and the European Environment

Agency thesaurus, GEMET. The participants were allowed to use the third thesaurus as background knowledge to match the other two for the construction of any of the three alignments.

7.1 Test set

The task of this case consists of matching three thesauri represented in SKOS:

GEMET: The European Environment Agency (EEA) GEneral Multilingual Environmental Thesaurus, version July 2007. This thesaurus consists of 5.298 concepts, each with descriptor terms in all of its 22 languages (bg, cs, da, de, el, en, en-US, es, et, eu, fi, fr, hu, it, nl, no, pl, pt, ru, sk, sl, sv).

AGROVOC: The United Nations Food and Agriculture Organization (FAO) AGROVOC thesaurus, version February 2007. This thesaurus consists of 28.445 descriptor terms, i.e., prefered terms, and 12.531 non-descriptor terms, i.e., alternative terms. AGROVOC is multilingual in eleven languages (en, fr, de, es, ar, zh, pt, cs, ja, th, sk).

NALT: The United States National Agricultural Library (NAL) Agricultural thesaurus, version 2007. This thesaurus consists of 42.326 descriptor terms and 25.985 non-descriptor terms. NALT is monolingual, English.

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

7.2 Results

The evaluation procedure used is the same as for the food task with the exception that we used slightly different categories of sample topics.

For the evaluation of precision we distinguished six 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.), geographical terms (countries, regions, etc.), health risk management (pollution, food, air, water, disasters, etc.), natural resources (fishery, forestry, agriculture, mining, etc.), and the remaining concepts (administration, materials, military aspects, etc.).

For the evaluation of recall we used a set of sub-hierarchies of the thesauri about: Geographical concepts like countries and place types (the Baltic states, alluvial plains, etc.), fishery (fishing equipment, aquaculture methods, etc.), and animal husbandry (animal diseases, animal housing, etc.).

8 Library

8.1 Test set

The National Library of the Netherlands (KB) maintains two large collections of books: the Deposit Collection, containing all the Dutch printed publications (one million

items), and the Scientific Collection, with about 1.4 million books mainly about the history, language and culture of the Netherlands.

Each collection is annotated using its own indexing system and controlled vocabulary. The Scientific Collection is described using the GTT thesaurus, a huge vocabulary containing 35.194 general concepts ranging from Wolkenkrabbers (Sky-scrapers) to Verzorging (Care). The books in the Deposit Collection are mainly indexed against the Brinkman thesaurus, which contains a large set of headings (5.221) for describing the overall subjects of books. Both thesauri have similar coverage (2.895 concepts actually have exactly the same label) but differ in granularity.

For each concept in the two thesauri, the usual detailed lexical information is provided: preferred labels (each concept has exactly one of them), synonyms (961 for Brinkman, 14.607 for GTT), extra hidden labels (134 for Brinkman, a couple of thousands for GTT) or scope notes (6236 for GTT, 192 for Brinkman). The language of both thesauri is Dutch[6], which makes this track ideal for testing matching systems in a non-English situation.

The two thesauri also provide structural information for their concepts, in the form of *broader* and *related* links. However, GTT contains only 15.746 hierarchical *broader* links between 35.194 concepts and 6.980 associative *related* links. Within the Brinkman thesaurus, there are 4.572 hierarchical links and 1.855 associative ones. On average, one can expect at most one parent per concept, for an average depth of 1 and 2, respectively[7]. The structural information found in the case is very poor.

For the purpose of matching, the two thesauri are represented in SKOS format. In case the participants' tool cannot process the SKOS data, OWL versions were provided, according to the conversion rules detailed on the track page[8].

8.2 Evaluation and results

Three participants produced their alignments:

- Falcon: 3.697 `exactMatch` correspondences;
- DSSim: 9.467 `exactMatch` correspondences;
- Silas: 3.476 `exactMatch` correspondences and 10.391 `relatedMatch` correspondences.

Two evaluation procedures were chosen, each of them motivated by a potential case of alignment usage. The first one is *thesaurus merging*, where the alignment is used to build a new, unified thesaurus from GTT and Brinkman thesauri. Evaluation in such a context requires assessing the validity of each individual alignment, which leads to a "standard" alignment evaluation procedure, using `exactMatch` and `relatedMatch` links. The second usage scenario for the alignment is *annotation translation* from one thesaurus to the other. Here, books are annotated using one thesaurus, and the alignment is used to produce a corresponding annotation using the other thesaurus. This scenario is particularly useful when one of the two thesaurus will be

[6] A quite substantial part of GTT concepts (around 60%) also have English labels.
[7] Particularly, the GTT thesaurus has 19.752 root concepts.
[8] `http://oaei.ontologymatching.org/skos2owl.html"`

dropped, and a massive amount of legacy data has to be converted to the remaining annotation system.

Evaluation in a thesaurus merging scenario

For this evaluation task, there was no reference alignment available. Given the size of the vocabularies, it was impossible to build one, so that we performed a manual evaluation on participants' results.

Inspired by the anatomy and food tracks of last year's OAEI, we opted for evaluating precision using a reference alignment we computed based on a lexical procedure[9] that gives 3.659 reliable equivalence links. We were also able to produce quantitative measures for coverage, which we define here as the proportion of all correct correspondences found in an alignment divided by the total number of correct correspondences produced by all participants and those in the reference. Coverage is proportional to recall and also provides an upper bound for it.

For manual evaluation, the set of all *equivalence* correspondences[10] was partitioned into parts unique to each combination of participant alignments plus reference set (15 parts in all). For each of those parts which was not in the lexical reference alignment, a sample of correspondences was selected, and evaluated manually. A total of 330 correspondences were assessed by two Dutch native experts.

>From the assessments of these parts, precision and coverage were calculated with their 95% confidence intervals, taking the sampling size and evaluator variability into account. The results are shown in Table 6 which shows clearly that Falcon performs better than both other participants, Silas comes second.

Alignment	Precision	Coverage
DSSim	0.134 $\pm$ 0.019	0.31 $\pm$ 0.19
Silas	0.786 $\pm$ 0.044	0.661 $\pm$ 0.094
Falcon	0.9725 $\pm$ 0.0033	0.870 $\pm$ 0.065

Table 6. Comparison of precision and coverage for the thesaurus merging scenario.

A detailed analysis reveals that Falcon results are very close to the lexical reference[11] which explains their observed quality. DSSim also uses lexical comparisons, but its edit-distance-like approach is more prone to error: an estimated between 20 and 200

[9] This makes use of direct comparison between concept labels, but also using a Dutch morphology database that allows to recognize grammatical variants of a word, e.g., between singular or plural forms.

[10] We did not proceed with manual evaluation of the *related* links, as only one participant provided such links, and their manual assessment is much more error-prone.

[11] 3.493 links are common to Falcon and the reference, Falcon has 204 correspondences not in the reference of which 100 are good and the lexical reference has 166 correspondences not in Falcon.

out its 8.399 correspondences not in the lexical reference are correct[12]. The Silas system is the one which succeeds most in adding to the lexical reference: 234 of its 976 "non-lexical" correspondences are correct, but Silas failed to reproduce one third of the lexical reference correspondences, therefore his coverage is relatively low.

Evaluation in an annotation translation scenario

Out of KB 2.4 million books, 250.000 actually belong both to KB Scientific and Deposit collections, and are therefore already indexed against both GTT and Brinkman thesauri. Here, the existing Brinkman indices from this dually indexed collection are taken as a reference alignment that an annotation translation system must aim to match. That is, for each book in the given corpus, we compare its existing (manually constructed) Brinkman index with the one computed from the GTT-Brinkman alignment produced by participants.

Evaluation settings. The correspondences sent by participants are transformed into mapping rules, i.e.,

$$R : g_r \rightarrow B_r,$$

where g_r is one GTT term and B_r is a set of Brinkman terms, to which the GTT term is matched. Note that $|B_r| \geq 1$, because some GTT terms are matched to more than one Brinkman term.

The data set consists of all dually indexed books, 243.887 in total. Each book has both GTT and Brinkman annotations, denoted as G_t and B_t. The real GTT annotation G_t is used to fire the transformed mapping rules.

If the GTT term of one rule is contained by the GTT annotation of one book, i.e., $g_r \in G_t$, then this book is *fired* by this rule. As one book can be fired by multiple rules, the union of the right parts of these rules forms the translated Brinkman annotation of this book, denoted as $B_r^{'}$. If this set of translated Brinkman terms overlaps the real Brinkman annotation of this book, i.e., $B_t \cap B_r^{'} \neq 0$, we consider this book as *matched*.

Evaluation measures.

– At the book level, we measure how many books are fired by these rules and how many of them are actually matched, i.e.,

$$P_b = \frac{\#books\_matched}{\#books\_fired}, \quad R_b = \frac{\#books\_matched}{\#all\_books},$$

where $\#all\_books$ is the number of the whole dually indexed books, $\#books\_fired$ is the number of fired books and $\#books\_matched$ is the number of matched books.

[12] Out of the selection of 86 correspondences in the set of 8.363 correspondences unique to DSSim not a single one was evaluated as correct by the human evaluators.

– At the annotation level, we measure how many translated terms are correct, how many real Brinkman annotation terms are missed and the combined measure of these two, i.e.,

$$P_a = \frac{\sum \frac{\#correct}{|B_r{}'|}}{\#books\_fired}, \quad R_a = \frac{\sum \frac{\#correct}{|B_t|}}{\#all\_books}, \quad J_a = \frac{\sum \frac{\#correct}{|B_t \cup B_r{}'|}}{\#all\_books}$$

where $\#correct$ is the number of the translated Brinkman terms which are actually used for the book.

The ultimate measure for alignment quality here is at the annotation level. The Jaccard overlap measure between found concepts and correct ones, i.e., J_a, plays a similar role as the F-measure does in information retrieval. Measures at the book level to some extent indicate users' (dis)satisfaction with the built system. A R_b of 60% means that the alignment does not produce any useful candidate for the rest 40% of the books.

We would like to mention that averaging these numbers naturally reflect the importance of thesaurus concepts: a good translation rule for a frequently used concept is more important than than that for a concept which is rarely used. We opted for this *micro-average* stance because it suits the real application context better.

Evaluation results. Table 7 gives an overview of the evaluation results when we only use the `exactMatch` correspondences. Falcon and Silas perform similarly, and much ahead of DSSim. As shown Figure 3, on the one hand, nearly half of the books are fired by at least one rule, and more than 65% of them have at least one Brinkman term translated correctly. On the other hand, at the annotation level, the translated results are not ideal: nearly half of the translated terms are not really used and more than 60% of the real Brinkman annotation is not found. We already pointed out that the correspondences from Falcon are mostly generated by lexical similarity. This indicates that lexical equivalent correspondences are not the only solution to the annotation translation scenario, which confirms the sensitivity of alignment evaluation methods to certain application scenarios.

Correspondences which are classified as `relatedMatch` are also useful in the annotation translation scenario. However, among three participants, only Silas generated `relatedMatch` correspondences. We combined `relatedMatch` correspondences with the `exactMatch` ones and transformed them into 8.410 rules. As shown in Figure 4, the use of `relatedMatch` correspondences increase the chances of a book being fired, also the recall of the translated set of Brinkman terms. However, as expected, the precision decreases because some noisy results are introduced at the same time.

8.3 Discussion

The first comment on this track concerns the *form* of the alignment returned by the participants, especially *wrt.* the type and cardinality of alignments.

First, all three participants proposed alignments using the SKOS links we asked for. However, only symmetric links (`exactMatch` and `relatedMatch`) were used:

Participant	$\#rules$	$\#books\_fired$	P_b	R_b	P_a	R_a	J_a
Falcon	3.618	183.754	65.32%	49.21%	52.63%	36.69%	30.76%
Silas	3.208	175.309	66.05%	47.48%	53.00%	35.12%	29.22%
DSSim	9.467	188.165	18.59%	14.34%	13.41%	9.43%	7.54%

Table 7. Performance of `exactMatch` correspondences produced by three participants.

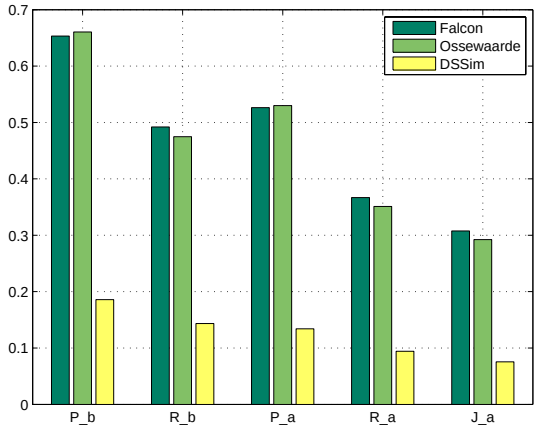

Fig. 3. Comparison between all `exactMatch` correspondences generated by Falcon, Silas and DSSim.

no participants proposed hierarchical `broader` and `narrower` links. Yet these links are useful for the application scenarios at hand. The `broader` links are useful to attach concepts which cannot be matched to an equivalent corresponding concept but a more generic or specific one, as the two thesauri have different granularity but the same general scope.

Second, there is no precise handling of one-to-many or many-to-many alignments. Sometimes a concept from one thesaurus is matched to several concepts from the other. This proves to be very useful, especially in the annotation translation scenario where concepts attached to a book should ideally be translated as a whole. As a result, we have to post-process alignments, building multi-concept correspondences from alignment that initially do not contain such links. This processing makes the evaluation of the relative quality of the alignments more difficult for the annotation scenario.

Of course these problems can be anticipated by making participants more aware of the different scenarios that are going to guide the evaluation. The campaign's timing made it impossible this year, but this is an option we would like to propose for next campaigns.

The results we have obtained also show that the performance of matchers vary from one scenario to the other, highlighting the strengths of different approaches. For the

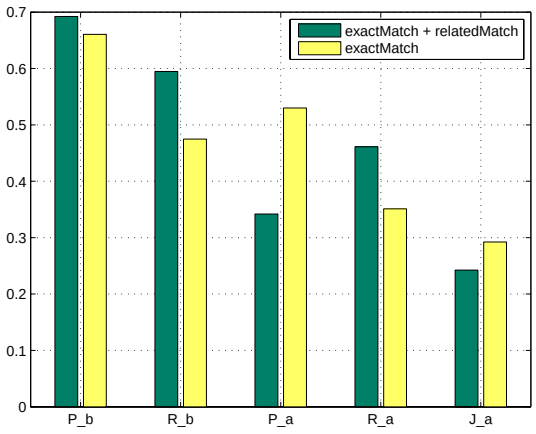

Fig. 4. The influence of `relatedMatch` correspondences in the case of Silas.

merging scenario, Falcon outperforms the two other participants. While in the translation scenario, Silas detects links based on extensional information of concepts[13], performs similarly as Falcon does.

Finally, we would like to discuss the overall quality of the results. The annotation translation scenario showed a maximum precision of 50%, and around 35% for recall. This is not much, but we have to take account of the fact that this scenario involves a high degree of variability: different annotators may choose different subject headings for a same book. Future manual evaluations[14], compensating for variability, could therefore give a better view of the quality.

This still leaves the low coverage of alignments with respect to the thesauri, especially GTT: in the best case, only 9.500 of its 35.000 concepts were linked to some Brinkman concept. This track, arguably because of its Dutch language context, seems to be difficult. The Silas system results are partially based on real book annotations show that the case can benefit from the release of such extensional information. We will investigate this option for future campaigns.

9 Conference

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

[13] It is important to mention here that Silas was actually applied on a set of books which is not the one we used for evaluation.

[14] A manual evaluation of the annotations by KB experts is in progress. We however had no time to complete this in time for this report.

9.1 Test set

The collection consists of fourteen ontologies in the domain of organizing conferences. Ontologies have been developed within the OntoFarm project[15]. In contrast to the last year's conference track, there are four new ontologies. Original ten ontologies have been slightly adjusted. 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 alignment among their entities 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 number of classes, properties, their DL expressivity, but also in underlying resources. Nine ontologies are based on *tools* supporting the task of organizing conferences, two are based on experience of people with *personal participation* in a conference organization, and three are based on *web pages* of concrete conferences.

Participants were expected to provide either complete alignments or interesting correspondences (nuggets), for all or some pairs of ontologies. There is no reference alignment. Instead, organizers of this track offer a posteriori evaluation of results in part manually and in part by data-mining techniques. Manual evaluation will produce statistics such as precision and will also serve as input into data-mining based evaluation. During manual evaluation some interesting correspondences will be chosen as a background material for the consensus building discussion.

9.2 Results

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. Let us make some observations:

- The ASMOV team and the Falcon team delivered totally 91 alignments. All ontologies were matched to each other. The Lily team also matched all ontologies to each other, moreover they also matched ontologies to themselves. The OLA2 team and the OntoDNA team matched all ontologies to each other, including separately each direction of correspondences.
- In order to make evaluation process more balanced, we transformed all results of participants into 91 alignments, except results of the SEMA tool. They delivered 13 alignments by matching all ontologies to the EKAW ontology.
- Four participants delivered correspondences with certainty factors between 0 and 1 (Falcon, OLA2, OntoDNA, and SEMA); two only (ASMOV and Lily) delivered "certain" correspondences.

[15] `http://nb.vse.cz/~svatek/ontofarm.html`

– As well as last year only equivalence (e.g., no subsumption) relations were delivered.
– Three participants delivered class-to-class correspondences, property-to-property correspondences as well as property-to-class correspondences (ASMOV, Lily, and OLA2); two participants (Falcon and OntoDNA) delivered class-to-class correspondences and property-to-property correspondences; the SEMA team delivered class-to-class correspondences.

Results of further analysis will be soon available on the Conference track web page[16]; these results will then serve as starting material for the consensus workshop discussion.

10 Lesson learned and suggestions

The most important applied lesson learned from last year is that we have been able to revise the schedule so we had more time for evaluation. But there remain lessons not really taken into account that we identify with an asterisk (*). So we reiterate those lessons that still apply with new ones, including:

A) This is a trend that there are now more matching systems and more systems are able to enter such an evaluation. This is very encouraging for the progress of the field.
B*) We also see systems that enter the campaign for several times. This means that we are not dealing with a continuous flow of prototypes but with systems on which there is a persistent development. These systems tend to improve over years.
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 seem 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. We would also like to have more test cases for expressive ontologies.
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 features one by one. This has still not been done this year but we are considering it seriously for the next year.
F) We have detected this year, through some random verifications, some submissions which were not strictly complying to the evaluation rules. We may have to be more strict about control in future.
G) Contrary to what has been noted in 2006, a significant number of systems were unable to output syntactically correct results (i.e., automatically usable by another program). Since fixing these mistakes by hand is becoming too much work, we plan to go towards automatic evaluation in which participants have to input correct results.

[16] http://nb.vse.cz/~svabo/oaei2007/#eval

H) There seems to be partitions of the systems: between systems able to deal with large test sets and systems unable to do it, between system robust on all tracks and those which are specialized (see Table 2). These observations remain to be further analyzed.

11 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, especially expressive ontologies;
- 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);
- Developing a definition of test hardness.

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

12 Conclusion

This year we had more systems that entered the evaluation campaign as well as more systems managed to produce better quality results compared to the previous years. Each individual test case had more participants than ever. This shows that, as expected, the field of ontology matching is getting stronger (and we hope that evaluation has been contributing to this progress).

On the side of participants, it seems that there is clearly a problem of size of input that should be addressed in a general way. We would like to see more participation on the large test cases. On the side of organizers, each year the evaluating of matching systems becomes more complex.

Most of the participants have provided description of their systems and their experience in the evaluation[17]. These OAEI papers, like the present one, have not been peer reviewed. 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`.

[17] The SCARLET system is described in [11].

Acknowledgments

We warmly thank each participant of this campaign. 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 the following up OAEI papers.

We are grateful to Lourens van der Meij and Shenghui Wang who have made crucial contributions to implementation and reporting on the Library track. The following persons were involved in the Food and Environment tracks: Lori Finch (National Agricultural Library, US department of agriculture); Johannes Keizer, Margherita Sini, Gudrun Johannsen, Patricia Merrikin (Food and Agriculture Organization of the United Nations); Jan Top, Nicole Koenderink, Lars Hulzebos, Hajo Rijgersberg, Keen-Mun de Deugd (Wageningen UR); Fred van de Brug (TNO Quality of Life) and Evangelos Alexopoulos (Unilever). We would like to thank the teams of Agricultural Organization of the United Nations (FAO) for allowing us to use their ontologies.

We are grateful to Martin Ringwald and Terry Hayamizu for providing the reference alignment for the anatomy ontologies.

We also thank the other members of the Ontology Alignment Evaluation Initiative Steering committee: Wayne Bethea (John Hopkins University, USA), Alfio Ferrara (Università degli Studi di Milano, Italy), Lewis Hart (AT&T, USA), Tadashi Hoshiai (Fujitsu, Japan), Todd Hughes (DARPA, USA), Yannis Kalfoglou (University of Southampton, UK), John Li (Teknowledge, USA), Miklos Nagy (The Open University (UK), Natasha Noy (Stanford University, USA), Yuzhong Qu (Southeast University (China), York Sure (University of Karlsruhe, Germany), Jie Tang (Tsinghua University (China), Raphaël Troncy (CWI, Amsterdam, The Netherlands), Petko Valtchev (Université du Québec à Montréal, Canada), and George Vouros (University of the Aegean, Greece).

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.12/06 “Integration of approaches to ontological engineering: design patterns, mapping and mining” and no.20/07 “Combination and comparison of ontology mapping methods and systems”.

References

1. Zharko Aleksovski, Warner ten Kate, and Frank van Harmelen. Exploiting the structure of background knowledge used in ontology matching. In *Proceedings of the ISWC workshop on Ontology Matching*, pages 13–24, Athens (GA US), 2006.
2. Ben Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Proceedings of the K-Cap 2005 workshop on Integrating ontologies*, Banff (CA), 2005.
3. Paolo Avesani, Fausto Giunchiglia, and Mikalai Yatskevich. A large scale taxonomy mapping evaluation. In *Proceedings of the 4th International Semantic Web Conference (ISWC)*, pages 67–81, Galway (IE), 2005.
4. Oliver Bodenreider, Terry F. Hayamizu, Martin Ringwald, Sherri De Coronado, and Songmao Zhang. Of mice and men: Aligning mouse and human anatomies. In *Proceedings of the American Medical Informatics Association (AIMA) Annual Symposium*, pages 61–65, 2005.

5. Marc Ehrig and Jérôme Euzenat. Relaxed precision and recall for ontology matching. In *Proceedings of the K-Cap 2005 workshop on Integrating Ontologies*, pages 25–32, Banff (CA), 2005.
6. Jérôme Euzenat. An API for ontology alignment. In *Proceedings of the 3rd International Semantic Web Conference (ISWC)*, pages 698–712, Hiroshima (JP), 2004.
7. Jérôme Euzenat, Malgorzata Mochol, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Svab, Vojtech Svatek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative 2006. In Pavel Shvaiko, Jérôme Euzenat, Natalya Noy, Heiner Stuckenschmidt, Richard Benjamins, and Michael Uschold, editors, *Proceedings of the ISWC workshop on Ontology Matching, Athens (GA US)*, pages 73–95, 2006.
8. Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. Discovering missing background knowledge in ontology matching. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, pages 382–386, Riva del Garda (IT), 2006.
9. Fausto Giunchiglia, Mikalai Yatskevich, and Paolo Avesani. A large scale dataset for the evaluation of matching systems. In *Posters of the 4th European Semantic Web Conference (ESWC)*, Innsbruck (AU), 2007.
10. Marta Sabou, Mathieu d'Aquin, and Enrico Motta. Using the semantic web as background knowledge for ontology mapping. In *Proceedings of the ISWC workshop on Ontology Matching*, pages 1–12, Athens (GA US), 2006.
11. Marta Sabou, Jorge Gracia, Sophia Angeletou, Matthieu d'Aquin, and Enrico Motta. Evaluating the semantic web: A task-based approach. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, Busan (KR), 2007.
12. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the ISWC workshop on Evaluation of Ontology-based tools (EON)*, Hiroshima (JP), 2004.

Grenoble, Amsterdam, Trento, Mannheim, and Prague, November 11th, 2007