HAL
open science

Results of the Ontology Alignment Evaluation Initiative 2007

Jérôme Euzenat, Antoine Isaac, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb, Vojtech Svátek, Willem Robert van Hage, Mikalai Yatskevich

To cite this version:

Jérôme Euzenat, Antoine Isaac, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, et al.. Results of the Ontology Alignment Evaluation Initiative 2007. Proc. 2nd ISWC 2007 international workshop on ontology matching (OM), Nov 2007, Busan, South Korea. pp.96-132. hal-00822893

HAL Id: hal-00822893

https://inria.hal.science/hal-00822893v1

Submitted on 15 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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,aisaac}@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. Ontology matching consists of finding correspondences between ontology entities. OAEI campaigns aim at comparing ontology matching systems on precisely defined test sets. Test sets can use ontologies of different nature (from expressive OWL ontologies to simple directories) and use different modalities (e.g., blind evaluation, open evaluation, consensus). OAEI-2007 builds over previous campaigns by having 4 tracks with 7 test sets followed by 18 participants. This is a major increase in the number of participants compared to the previous years. Moreover, 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 paper improves on the "First results" initially published in the on-site proceedings of the ISWC+ASWC workshop on Ontology Matching (OM-2007). The only official results of the campaign, however, are on the OAEI web site.

[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) [13]. 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 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 ontology matching. 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 on its execution.

2.1 Tracks and test cases

This year's campaign has consisted of four tracks gathering seven data sets 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 the Open directory or Yahoo's). It has more than four thousand 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. In addition, we run application dependent evaluation.

The conference track and consensus workshop (§9): Participants were 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 was no a priori reference alignment. Organizers of this track offered manual a posteriori evaluation of results. For a selected sample of correspondences, consensus was sought at the workshop and the process of its reaching was tracked and 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, OWL	narrow-, exact-, broadMatch	1	blind+external
environment	SKOS, OWL	narrow-, exact-, broadMatch	1	blind+external
library	SKOS, OWL	narrow-, exact-, broad-, relatedMatch	1	blind+external
conference	OWL-DL	=, $\leq$	1	blind+consensual

Table 1. Characteristics of test cases (open evaluation is made with already published reference alignments, 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 as well as measures compensating for the lack of complete reference alignments.

In addition, the Library test case featured an application-specific evaluation and a consensus workshop has been held for evaluating particular correspondences.

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 18 in 2007. We can also observe a common trend: participants who keep on developing their systems improve the evaluation results over time.

We have had not enough time 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 others. 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.

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

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							√	
SODA	√	√						
TaxoMap	√	√	√					
X-SOM	√	√	√	√	√			
Total=18	10	14	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.

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 ontologies are described in OWL-DL and serialized in the RDF/XML format. The reference ontology is that of test #101. It contains 33 named classes, 24 object properties, 40 data properties, 56 named individuals and 20 anonymous individuals. Participants have to match this reference ontology with the variations. Variations are focused on the characterization of the behavior of the tools rather than having them compete on real-life problems. They are organized in three groups:

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) 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) found on the web and left mostly untouched (there were added xmlns and xml:base attributes).

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 test 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.

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		SODA		TaxoMap		X-SOM	
test	Prec.	Rec.	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	0.67	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	0.96	0.50	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	0.51	0.41	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	0.51	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		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.

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 scores are improved and the six leading systems are closer to each others). This indicates that the close followers of the 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 from 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. In addition, 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 its precision and decreased its 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.

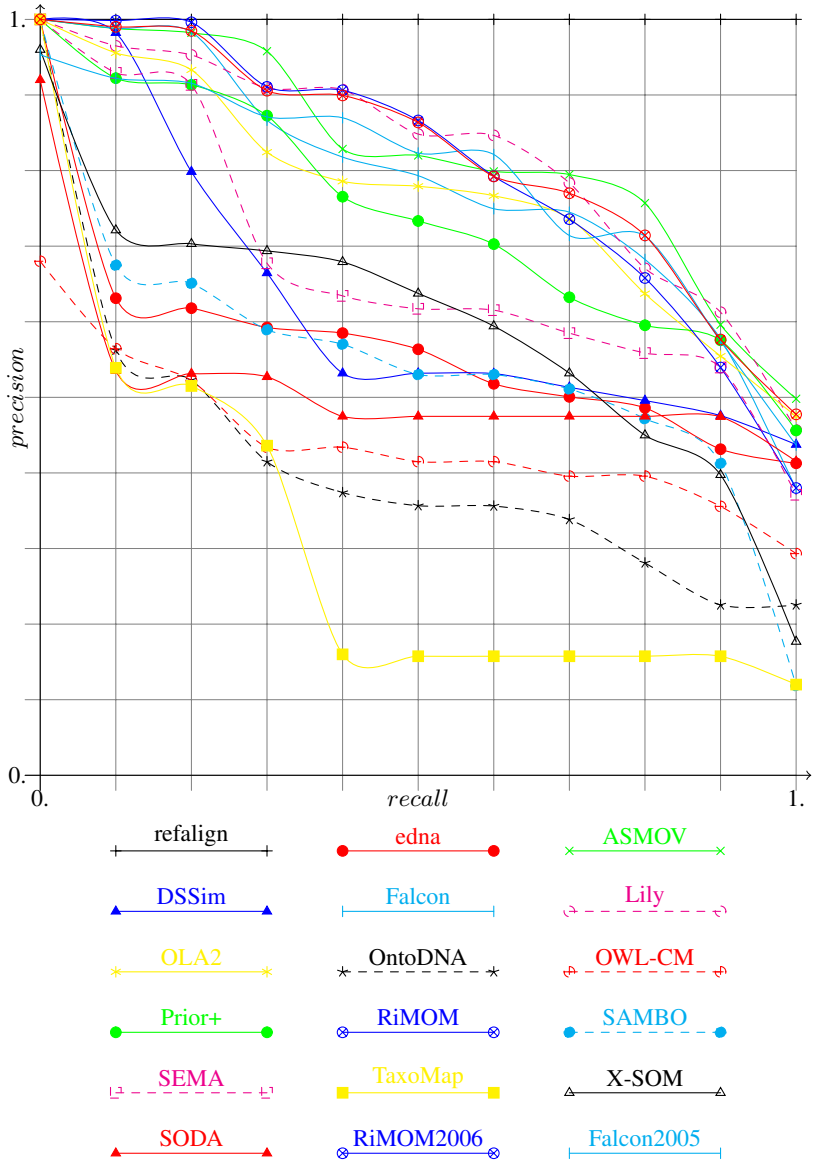

Fig. 1. Precision/recall graphs. They cut the results given by the participants under a threshold necessary for achieving $n\%$ recall and compute the corresponding precision. Systems for which these graphs are not meaningful (because they did not provide graded confidence values) are drawn in dashed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

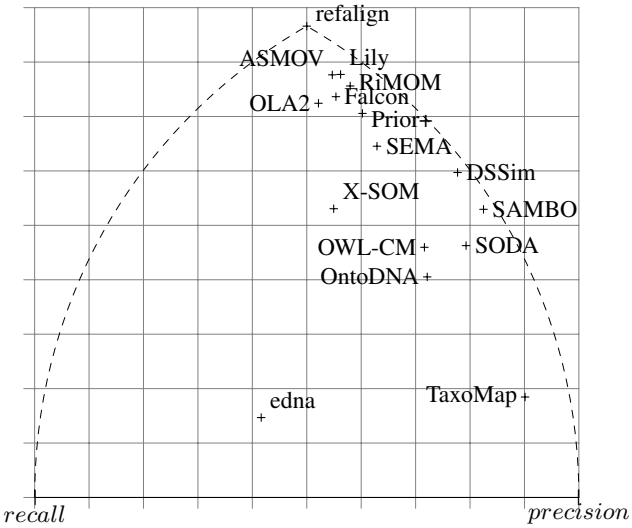

Fig. 2. Each point expresses the position of a system with regard to precision and recall.

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 technologies 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 a reference alignment R and an alignment S generated by the naive string equality matching,

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

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

$recall+$ is defined as follows:

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

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 fall 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 ASMOV 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 task #1 compared to almost all matching systems of type C. Only TaxoMap and AgreementMaker generate alignments with higher recall but a significant

[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	Task #1				Task #2		Task #3		Recall+	
		Runtime	Prec	Rec	F-meas	Prec	Rec	Prec	Rec	#1	#3
AOAS	A	2h	0.928	0.804	0.861	-	-	-	-	0.505	-
SAMBO	A	6 h	0.845	0.786	0.815	-	-	-	-	0.580	-
ASMOV	B	15 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 task #1. The values for recall+ are presented in the rightmost columns for task #1 and #3.

loss in precision. We would have expected the participating systems 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 task #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; 11]. 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 task #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 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

In OAEI-2007, 9 out of 18 matching systems participated on the web directories data set, while in OAEI-2006, 7 out of 10, and in OAEI-2005, 7 out of 7 did it. Only the Falcon system participated in all three evaluations of the web directories data set. In 2007, participating systems demonstrated substantially higher quality results than in previous two years.

Precision, recall and F-measure of the systems on the web directories test case are shown in Figure 3. These indicators have been computed following the TaxMe and TaxMe2 methodologies [3; 9] and with the help of Alignment API [6].

Let us make several observations concerning quality of the results of the participated systems. In particular, the average F-measure of the systems increased from approximately 29% in 2006 to 49% in 2007. The highest F-measure of 71% was demonstrated by the OLA2 system in 2007. The average precision of the systems increased from approximately 35% in 2006 to 57% in 2007. The highest precision of 62% was demonstrated by both the OLA2 system and X-SOM in 2007. The average recall of the systems increased from approximately 22% in 2005 to 26% in 2006 and to 50% in 2007. The highest recall of 84% was demonstrated by the OLA2 system in 2007. Notice that in 2005 this data set allowed for estimating only recall, therefore in the above observations there are no values of precision and F-measure for 2005.

A comparison of the results in 2006 and 2007 for the top-3 systems of each of the years based on the highest values of the F-measure indicator is shown in Figure 4. The key observation is that quality of the best F-measure result of 2006 demonstrated by Falcon is almost doubled (increased by $\sim$1.7 times) in 2007 by OLA2. The best precision result of 2006 demonstrated by Falcon was increased by $\sim$1.5 times in 2007 by both OLA2 and X-SOM. Finally, for what concerns recall, the best result of 2005 demonstrated by OLA was increased by $\sim$1.4 times in 2006 by Falcon and further increased by $\sim$1.8 times in 2007 by OLA2. Thus, the OLA team managed to improve by $\sim$2.6 times its recall result of 2005 in 2007.

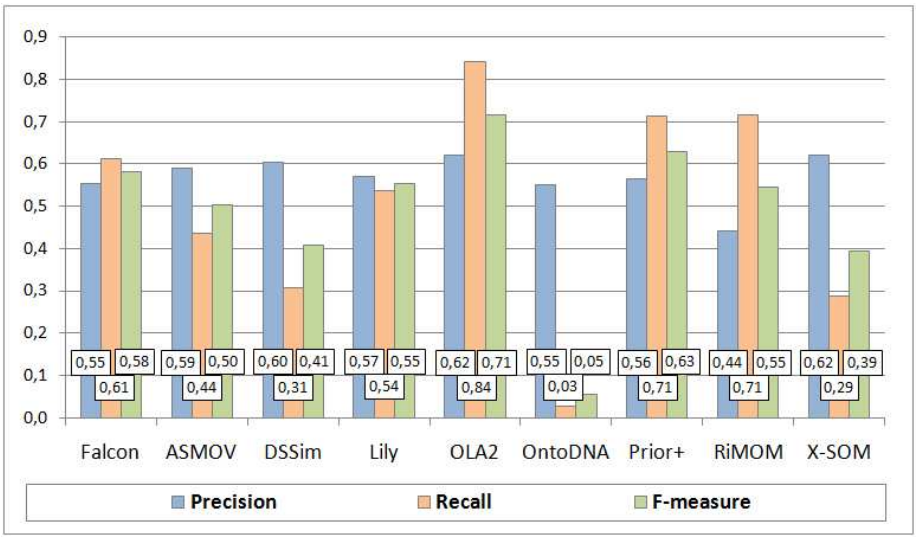

Fig. 3. Matching quality results.

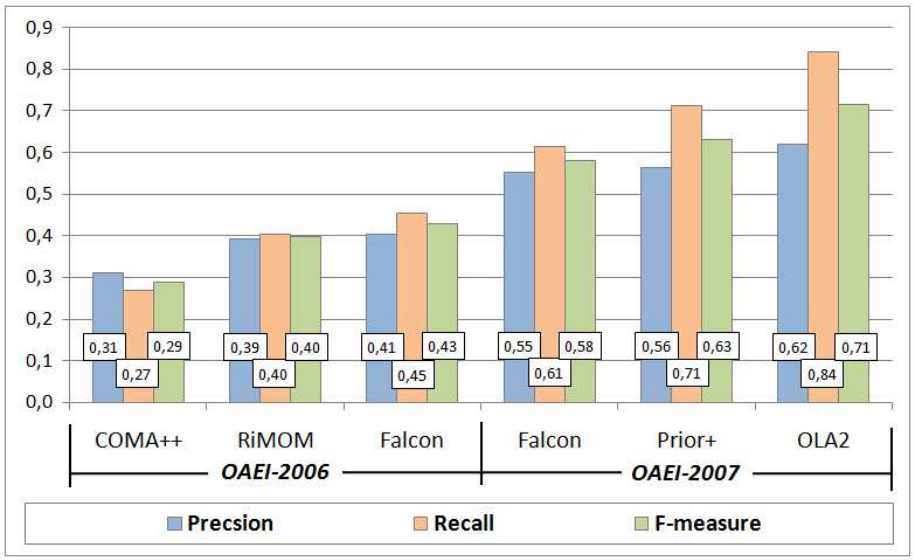

Fig. 4. Comparison of matching quality results in 2006 and 2007.

Partitions of positive and negative correspondences according to the system results are presented in Figure 5 and Figure 6, respectively.

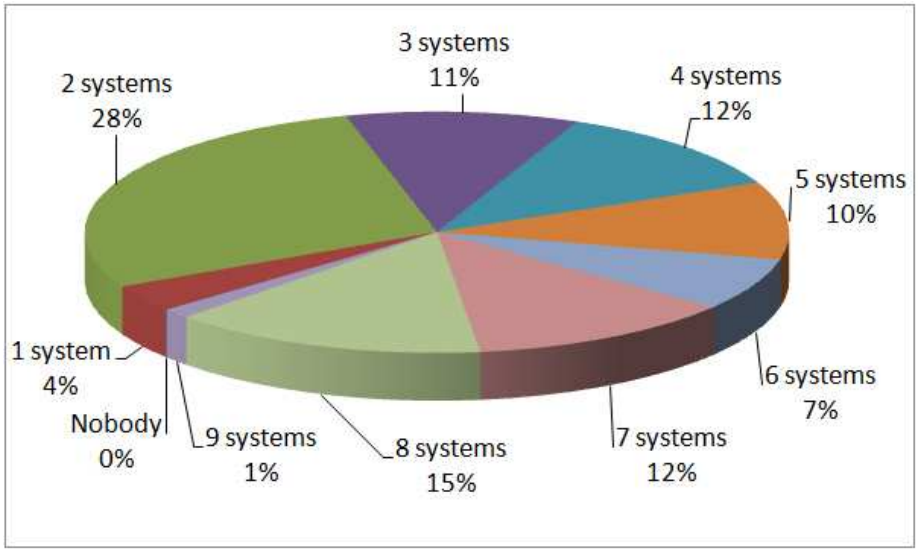

Fig. 5. Partition of the system results on positive correspondences.

Figure 5 shows that the systems managed to discover all the positive correspondences (Nobody - 0%). Only 15% of positive correspondences were found by almost all (8) matching systems. Figure 6 shows that almost all (8) systems found 11% of negative correspondences, i.e., mistakenly returned them as positive. The last two observations suggest that the discrimination ability of the data set is still high.

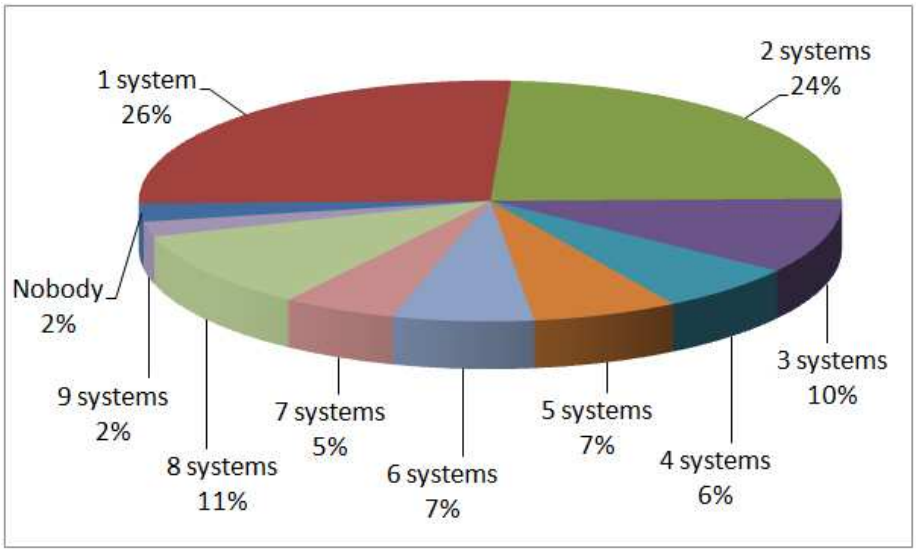

Fig. 6. Partition of the system results on negative correspondences.

Let us now compare partitions of the system results in 2006 and 2007 on positive and negative correspondences, see Figure 7 and Figure 8, respectively.

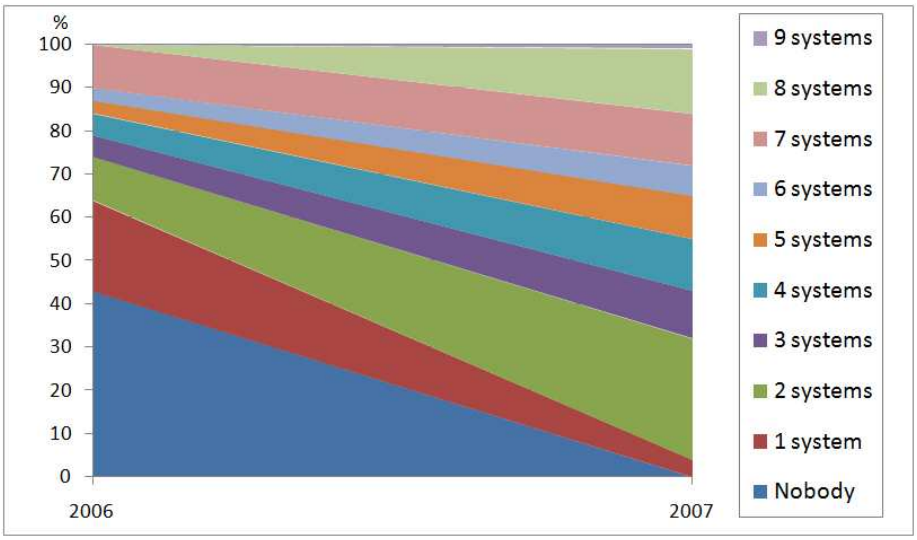

Fig. 7. Comparison of partitions of the system results on positive correspondences in 2006 and 2007.

Figure 7 shows that 43% of the positive correspondences have not been found by any of the matching systems in 2006, while in 2007 all the positive correspondences have been collectively found; see also how the selected regions (e.g., for 2 systems) consequently enlarge from 2006 to 2007.

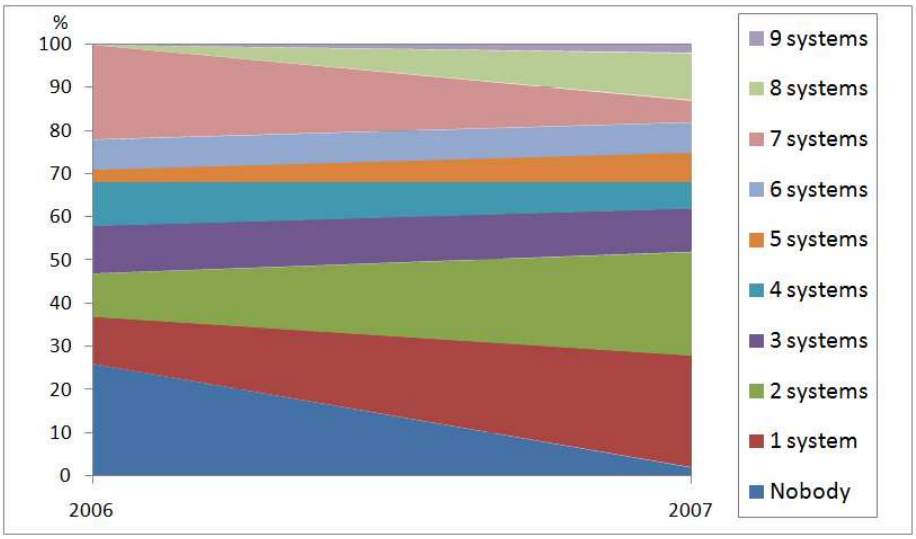

Fig. 8. Comparison of partitions of the system results on positive correspondences in 2006 and 2007.

Figure 8 shows that in 2006 in overall the systems have correctly not returned 26% of negative correspondences, while in 2007, this indicator decreased to 2%. In turn in 2006, 22% of negative correspondences were mistakenly found by all (7) the matching systems, while in 2007, this indicator decreased to 5%. An interpretation of these observations could be that systems keep trying various combinations of both "brave" and "cautious" strategies in discovering correspondences with a convergence towards better quality, since average precision increased from 2006 to 2007.

5.3 Comments

The key observation out of this evaluation is that the ontology matching community has made a substantial progress on the web directories task this year. In fact, as Figure 4 indicates, quality of the results is almost doubled from 2006 to 2007. This suggests that the systems experience fewer difficulties on the test case, although there still exists room for further improvements. Finally, as partitions of positive and negative correspondences indicate (see Figure 5 and Figure 6), the data set retains good discrimination ability, i.e., different sets of correspondences are still hard for the different systems.

6 Food

The food test case is another task in which the hierarchies come from thesauri, 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 formulated 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

Precision. In order to give dependable precision results within the time span of the campaign given a limited number of assessors we performed a combination of semi-automatic evaluation for alignments between taxonomical concepts and sample evaluation on roughly 5% of the other 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, i.e., the "hardness" or "complexity" of the alignment.

We distinguished four categories of topics in the thesauri that each required a different level of domain knowledge of the assessors: (i) taxonomical concepts (plants, animals, bacteria, etc.), (ii) biological and chemical terms (structure formulas, terms from genetics, etc.), (iii) geographical terms (countries, regions, etc.), and (iv) 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 rule works, because the taxonomical parts of the thesauri are based on the same sources. Samples from the other three categories were 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). The sizes of the categories and the part that was assessed are shown in Table 6.

topic	# alignments	# assessed alignments (sample size)
taxonomical	22.542	22.542
biological / chemical	3.816	200
geographical	1.284	86
miscellaneous	9.678	476

Table 6. Categories of alignments that were separately assessed for the estimation of precision.

Recall. To give dependable recall results within the time span of the campaign we estimated recall on a set of sample sub-hierarchies of the thesauri. Specifically, everything under the NALT concept animal health and all AGROVOC concepts that have alignments to these concepts and their sub-concepts, all oak trees (everything under the concept representing the Quercus genus), all rodents (everything under Rodentia), countries of Europe, and part of the geographical concepts below country level (cities, provinces, etc.). These sample reference alignments consisted of exactMatch, narrowMatch, and broadMatch alignments. The sizes of the samples are shown in Table 7, along with the percentage of exactMatch alignments in each sample.

topic	# alignments	% exactMatch
animal health	34	57%
oak trees (taxonomical)	41	84%
rodents (vernacular)	42	32%
Europe (country level)	74	93%
geography (below country level)	164	35%

Table 7. Reference alignments that were used for the estimation of recall.

Significance. As a significance test on the percentile scores of the systems we used the Bernoulli distribution. The performance (precision or recall) 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:

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

6.3 Results

Five participants took part in the OAEI-2007 food test case: South East University (Falcon-AO 0.7), 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 6.583 (X-SOM) alignments. This amounted to 37.384 unique alignments in total. Table 8 shows the total number of alignments that were submitted by each of the systems.

system	# alignments	alignment type
Falcon-AO	15.300	exactMatch
RiMOM	18.420	exactMatch
X-SOM	6.583	exactMatch
DSSim	14.962	exactMatch
SCARLET	81	exactMatch
	6.038	broadMatch & narrowMatch

Table 8. Number and type of alignments that were returned by the participating systems.

Best precision. The taxonomical parts of the thesauri accounted for by far the largest part of the alignments. The more difficult correspondences 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 alignments. This caused systems that did well at the taxonomical part to have a great advantage over the other systems. The Falcon-AO system performed consistently best at the largest two strata, taxonomical and miscellaneous, and thus achieved high precision. An overview of all the precision results is shown in Table 9. The results of the SCARLET system have been evaluated separately for each alignment type and hence are shown separately in Table 9.

Best recall. All systems except SCARLET only returned exactMatch alignments. This significantly limits recall. In Table 10 the first number represents recall of all types of alignment relations. Systems that only find exactMatch alignments are unable to achieve 1.00 here. The second number (between parentheses) shows recall of only exactMatch alignments. That means all systems can achieve 1.00 here. The RiMOM system had the highest recall for the OAEI 2006. This year, however, the Falcon-AO system has a higher recall than the RiMOM system. For some categories the RiMOM result equals to that of Falcon-AO, but on average the difference is significant.

Precision	Falcon-AO	RiMOM	X-SOM	DSSim	SCARLET e.M.	SCARLET b.M. & n.M.
taxonomical	0.81	0.54	0.26	0.37	0.60	0.13
bio/chem	0.91	0.84	0.92	0.86	1.00	0.17
geographical	0.95	0.97	1.00	0.94	0.00	1.00
miscellaneous	0.86	0.69	0.62	0.57	0.75	0.44
overall	**0.84**	**0.62**	**0.45**	**0.49**	**0.66**	**0.25**

Table 9. Precision results based on sample evaluation.

Recall	Falcon-AO	RiMOM	X-SOM	DSSim	SCARLET
animal health	0.21 (0.64)	0.21 (0.64)	0.00 (0.00)	0.06 (0.18)	0.00 (0.00)
oak trees	0.93 (1.00)	0.93 (1.00)	0.10 (0.12)	0.22 (0.24)	0.00 (0.00)
rodents	0.40 (0.71)	0.24 (0.42)	0.07 (0.10)	0.17 (0.29)	0.00 (0.00)
Europe	0.81 (0.97)	0.70 (0.84)	0.08 (0.10)	0.34 (0.40)	0.00 (0.00)
geography	0.32 (0.90)	0.26 (0.74)	0.05 (0.14)	0.18 (0.50)	0.01 (0.02)
overall	**0.49 (0.90)**	**0.42 (0.78)**	**0.06 (0.11)**	**0.20 (0.37)**	**0.00 (0.00)**

Table 10. Recall results based on sample evaluation. The numbers between parentheses show recall when only the exactMatch alignments of the reference alignments are considered.

7 Environment

The environment test set comprises three matching task 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 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 formulated 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 Evaluation procedure

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

Precision. For the evaluation of precision for the GEMET-AGROVOC and GEMET-NALT alignments we distinguished six categories of topics in the thesauri that each required a different level of domain knowledge of the assessors: (i) taxonomical concepts (plants, animals, bacteria, etc.), (ii) biological and chemical terms (structure formulas, terms from generics, etc.), (iii) geographical terms (countries, regions, etc.), (iv) natural resources (fishery, forestry, agriculture, mining, etc.), (v) health risk management (pollution, food, air, water, disasters, etc.), and (vi) the remaining concepts (administration, materials, military aspects, etc.). The results for the NALT-AGROVOC are shown in the section about the food task. The sizes of the categories and the part that was assessed are shown in Table 11.

	GEMET-AGROVOC		GEMET-NALT	
topic	# alignments	# assessed	# alignments	# assessed
taxonomical	500	39	802	33
biological / chemical	541	43	841	51
geographical	167	40	164	39
natural resources	412	51	450	39
health risk management	602	38	738	52
miscellaneous	1.884	48	1.988	51

Table 11. Categories of alignments that were separately assessed for the estimation of precision.

Recall. For the evaluation of recall we used a set of sub-hierarchies of the thesauri. Specifically, concepts from agriculture in the broad sense of the word, including: fishery (fishing equipment, aquaculture methods, etc.) and animal husbandry (animal diseases, animal housing, etc.), and geological concepts like countries and place types (the Baltic states, alluvial plains, etc.). The sizes of the samples are shown in Table 12, along with the percentage of exactMatch alignments in each sample.

	GEMET-AGROVOC		GEMET-NALT	
topic	# alignments	% exactMatch	# alignments	% exactMatch
agriculture	89	69%	92	66%
geology	136	64%	138	56%

Table 12. Reference alignments that were used for the estimation of recall.

7.3 Results

Two systems took part in the OAEI 2007 environment task: Falcon-AO 0.7 (South East University) and DSSim (Knowledge Media Institute). Both systems returned only exactMatch alignments. Table 13 shows the number of correspondences the two systems returned for each of the three tasks.

	# correspondences		
system	NALT-AGROVOC	GEMET-AGROVOC	GEMET-NALT
Falcon-AO	15.300	1.384	1.374
DSSim	14.962	3.030	4.278

Table 13. Number of correspondences that were returned by the participating systems.

Best precision. The GEMET thesaurus is very shallow compared to the AGROVOC and NALT thesauri, but it does offer definitions and labels in many languages. In consequence, lexical comparison is usually the only source of information that the matching system can exploit. This means that there is very little information for the matching systems to reason with. The Falcon-AO system performed best at both tasks, achieving a similar precision as with the easier NALT-AGROVOC task. An overview of all the precision results is shown in Table 14.

	GEMET-AGROVOC		GEMET-NALT	
Precision for	Falcon-AO	DSSim	Falcon-AO	DSSim
taxonomical	0.95	0.27	0.87	0.16
bio/chem	0.54	0.00	0.88	0.53
geographical	1.00	0.30	0.77	0.29
natural resources	1.00	0.53	0.95	0.32
health risk man.	0.95	0.38	0.88	0.50
miscellaneous	0.90	0.39	0.82	0.53
overall	**0.88**	**0.33**	**0.86**	**0.44**

Table 14. Precision results based on sample evaluation.

Best recall. The Falcon-AO system performs significantly better than the DSSim system on the GEMET-AGROVOC and GEMET-NALT tasks. However, it does not achieve similar recall scores as for the NALT-AGROVOC task.

	GEMET-AGROVOC		GEMET-NALT	
Recall for	Falcon-AO	DSSim	Falcon-AO	DSSim
agriculture	0.43 (0.62)	0.11 (0.16)	0.36 (0.54)	0.16 (0.25)
geology	0.37 (0.59)	0.18 (0.29)	0.26 (0.47)	0.17 (0.30)
overall	**0.39 (0.60)**	**0.15 (0.24)**	**0.30 (0.50)**	**0.16 (0.27)**

Table 15. Recall results based on sample evaluation. The numbers between parentheses show recall when only the exactMatch alignments of the reference alignments are considered.

8 Library

This is the last test case from the directory and thesauri track. It deals with two large Dutch thesauri.

8.1 Data 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.

Each collection is annotated – *indexed* – using its own 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 described 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.

Each concept has (exactly) one preferred label, synonyms (961 for Brinkman, 14.607 for GTT), extra hidden labels (134 for Brinkman, a couple of thousands for GTT) or scope notes (6.236 for GTT, 192 for Brinkman). The language of both thesauri is Dutch, albeit around 60% of GTT concepts also have English labels, which makes this track ideal for testing alignment in a non-English situation.

Concepts are also provided with structural information, in the form of *broader* and *related* links. However, GTT (resp. Brinkman) contains only 15.746 (resp 4.572) hierarchical *broader* links and 6.980 (resp. 1.855) associative *related* links. On average, one can expect at most one parent per concept, for an average depth of 1 and 2, respectively (in particular, the GTT thesaurus has 19.752 root concepts). The thesauri's structural information is thus very poor.

For the purpose of the OAEI campaign, the two thesauri were made available in the SKOS format. OWL versions were also provided, according to the – lossy – conversion rules detailed on the track page[6].

8.2 Evaluation and results

Three teams handed in final results: Falcon (3.697 `exactMatch` mappings), DSSim (9.467 `exactMatch` mappings), Silas (3.476 `exactMatch` mappings and 10.391 `relatedMatch` mappings). Two evaluation procedures were chosen, each of them motivated by a potential case of mapping usage.

Evaluation in a thesaurus merging scenario. The first scenario is *thesaurus merging*, where an 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 mapping, as in "standard" alignment evaluation.

[6] `http://oaei.ontologymatching.org/2007/library/`

Here, there was no reference alignment available. Given the size of the vocabularies, it was impossible to build one. Inspired by the anatomy and food tracks of OAEI 2006, we opted for evaluating precision using a reference alignment based on a lexical procedure. This makes use of direct comparison between labels, but also exploits a Dutch morphology database that allows to recognize variants of a word, e.g., singular and plural. 3.659 reliable equivalence links are obtained this way. We also measured coverage, which we define as the proportion of all good correspondences found by an alignment divided by the total number of good correspondences produced by all participants and those in the reference.

For manual evaluation, the set of all *equivalence* correspondences[7] was partitioned into parts unique to each combination of participant alignments plus reference set (15 parts in all). For each of those parts which were 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 these assessments, precision and coverage were calculated with their 95% confidence intervals, taking into account sampling size and evaluator variability. The results are shown in Table 16, which identifies clearly Falcon as performing better than both other participants.

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 16. Precision and coverage for the thesaurus merging scenario.

A detailed analysis reveals that Falcon results are very close to the lexical reference, which explains their observed quality. 3.493 links are common to Falcon and the reference, while Falcon has 204 correspondences not in the reference – of which 100 are good – and the lexical reference has 166 correspondences not identified by Falcon. DSSim also uses lexical comparisons, but its edit-distance-like approach is more prone to error: between 20 and 200 out its 8.399 correspondences not in the reference are correct. Silas is the one that succeeds most in adding to the reference: 234 of its 976 "non-lexical" correspondences are correct. But it fails to reproduce one third of the reference correspondences, therefore its coverage is relatively low.

Evaluation in an annotation translation scenario. The second usage scenario, aimed at indexers with an intricate expertise of Brinkman or GTT, consists in an *annotation translation* process supporting the re-indexing of GTT-indexed books with Brinkman concepts. This is particularly useful if GTT is dropped: a huge volume of legacy data has to be converted to the remaining annotation system.

This evaluation scenario requires building a tool that can interpret the correspondences provided by the different participants so as to translate existing GTT book annotations into equivalent Brinkman annotations. Based on the quality of the results for

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

books we know the correct annotations of, we can assess the quality of the initial correspondences. This approach, based on evaluation of user's information needs (here, book annotations) is more in line with the application-specific, end-to-end approach described in [14].

Evaluation settings and measures. The simple concept-to-concept correspondences sent by participants were transformed into more complex mapping rules that associate one GTT concept and a set of Brinkman concepts – some GTT concepts are indeed involved in several mapping statements. Considering `exactMatch` only, this gives 3.618 rules for Falcon, 3.208 rules for Silas and 9.467 rules for DSSim.

The set of GTT concepts attached to each book is then used to decide whether these rules are *fired* for this book. If the GTT concept of one rule is contained by the GTT annotation of a book, then the rule is fired. As several rules can be fired for a same book, the union of the consequents of these rules forms the translated Brinkman annotation of the book.

On a set of books selected for evaluation, the generated concepts for a book are then compared to the ones that are deemed as correct for this book. At the book level, we measure how many books have a rule fired on them, and how many of them are actually *matched* books, i.e., books for which the generated Brinkman annotation contains at least one correct concept. These two figures give a precision (P_b) and a recall (R_b) for this book level.

At the annotation level, we measure (i) how many translated concepts are correct over the annotation produced for the books on which rules were fired (P_a), (ii) how many correct Brinkman annotation concepts are found for all books in the evaluation set (R_a), and (iii) a combination of these two, namely a Jaccard overlap measure between the produced annotation (possibly empty) and the correct one (J_a).

The ultimate measure for alignment quality here is at the annotation level. Measures at the book level are used as a raw indicator of users' (dis)satisfaction with the built system. A R_b of 60% means that the alignment does not produce any useful candidate concept for 40% of the books. We would like to mention that, in these formulas, results are counted on a book and annotation basis, and not on a rule basis. This reflects the importance of different thesaurus concepts: a translation rule for a frequently used concept is more important than a rule for a rarely used concept. This option suits the application context better.

Automatic evaluation and results. Here, the reference set consists of 243.887 books belonging both to KB Scientific and Deposit collections, and therefore already indexed against both GTT and Brinkman. The existing Brinkman indices from these books are taken as a reference to which the results of annotation translation are automatically compared.

Table 17 gives an overview of the evaluation results when we only use the `exactMatch` mappings. Falcon and Silas perform similarly, and much ahead of DSSim. Nearly half of the books were given at least one correct Brinkman concept in the Falcon case, which corresponds to 65% of the books a rule was fired on. At the annotation level, half of the translated concepts are not validated, 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 lexically equivalent correspondences alone do not solve the annotation translation problem. It also confirms the sensitivity of mapping evaluation methods to certain application scenarios.

Participant	P_b	R_b	P_a	R_a	J_a
Falcon	65.32%	49.21%	52.63%	36.69%	30.76%
Silas	66.05%	47.48%	53.00%	35.12%	29.22%
DSSim	18.59%	14.34%	13.41%	9.43%	7.54%
Silas+related	69.23%	59.48%	34.20%	46.11%	24.24%

Table 17. Performance of annotation translations generated from correspondences.

Among the three participants, only Silas generated `relatedMatch` mappings. To evaluate their usefulness for annotation translation, we combined them with the `exactMatch` ones so as to generate a new set of 8,410 rules. As shown in the *Silas+related* line in Table 17, the use of `relatedMatch` mappings increases the chances of having a book given a correct annotation. However, unsurprisingly, precision of annotations decreases, because of the introduction of noisy results.

Manual evaluation and results. Automatic evaluation against existing annotations gives a first large and relatively cheap assessment of participants' results. Yet it is sensitive to *indexing variation*: several indexers annotating a same book, or a same annotator annotating it at different times, will select different concepts. We decided to perform an additional *manual* evaluation to assess the influence of this phenomenon, as well as to validate or invalidate the results of the automatic evaluation.

For this evaluation, we have partly followed the approach presented in [10]. First, a sample of 96 books was randomly selected among the dually annotated books annotated in 2006. On these books we applied the translation rules derived from each participants' results – using only the `exactMatch` links. For each book, the results of these different procedures are merged in a single list of candidate concept annotations. As we wanted some insight on the automatic evaluation based on existing Brinkman annotations, we also included these original annotations in the candidate lists.

To collect assessments of the candidate annotations, a paper form was created for each book in the sample. Each form constitutes an evaluation task where the evaluator validates the proposed annotations: for each of the candidates, she is asked whether it is *acceptable*[8] for an index. Afterwards, she is asked to select, among the candidates, the ones she would have *chosen as indices*. She also has the possibility to add to the list of chosen indices some concepts which are not in the proposed annotation. This form was validated by running a pilot evaluation.

The judges involved in the evaluation are four professional book indexers – native Dutch speakers – from the Depot department at the KB. Each of the evaluators assessed the candidates for every book in the test set.

[8] This precision is made to avoid too narrow choices, e.g., when the subject of the book is unclear, the thesaurus contains several concepts equally valid for the book, or when the evaluator feels other indexers could have selected indices different from hers.

Table 18 presents the acceptability assessments, averaged over the four evaluators. These are significantly and regularly higher than the figures obtained for automatic evaluation. This confirms the dependence of the scenario on the way indexing variability is taken into account in the evaluation setting.

Participant	P_a	R_a	J_a	P_a	R_a	J_a
Falcon	74.95%	46.40%	42.16%	52.63%	36.69%	30.76%
Silas	70.35%	39.85%	35.46%	53.00%	35.12%	29.22%
DSSim	21.04%	12.31%	10.10%	13.41%	9.43%	7.54%

Table 18. Comparison of correspondences as assessed by manual evaluation (left), and automatic evaluation results (right, from Table 17).

To assess *evaluation variability*, we computed the (Jaccard) overlap between the evaluators' assessments. On average, two evaluators agree on 60% of their assessments. We also measured the agreement between evaluators using Krippendorff's *alpha* coefficient – a common measure for computational linguistics tasks. The overall *alpha* coefficient is 0.62, which, according to standards, indicates a great variability. This is however to be put into perspective: the tasks usually analyzed with this coefficient, e.g., part-of-speech tagging, are less "variable" than subject indexing.

Indexing variability was first measured by assessing the original Brinkman indices for the books, which we had added in the candidate concepts to be evaluated. These concepts are the results of a careful selection, and do not render all the acceptable concepts for a book. It is therefore no surprise that the recall is relatively low (R_a =66.69%). However, it is very surprising to see that almost one original index concept out of five is not acceptable (P_a =81.60%). This result shows indeed that indexing variability matters a lot, even when the annotation selection criteria are made less selective.

To measure agreement between the indexers involved in our evaluation, we have computed the average Jaccard overlap between their chosen indices, as well as their Kripendorff's alpha. Again, we have quite a low overall agreement value – 57% for Jaccard, 0.59 for Krippendorff – which confirms the high intrinsic variability of the indexing task.

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. All three participants proposed alignments using the SKOS links we asked for. However, only symmetric links (`exactMatch` and `relatedMatch`) were used: no participant 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 mapped to an equivalent corresponding concept but a more general or specific one. This is likely to happen, since the two thesauri have different granularity but a 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 mapped 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 alignment results, building multi-concept correspondences from alignments which 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 will 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 matching systems vary from one scenario to the other, highlighting the strengths of different approaches. For the merging scenario, Falcon outperforms the two other participants. While in the translation scenario, Silas, which detects links based on extensional information of concepts[9], performs similarly to Falcon.

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 consider that this scenario involves a high degree of variability: different annotators may choose different concepts for a same book. The manual evaluation by KB expert illustrate this phenomenon, and show that under specific but realistic application conditions the quality of participant's result is more satisfactory.

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. Silas' results, which are partly based on real book annotations, demonstrate that the task can benefit from the release of such extensional information. We will investigate this option for future campaigns.

9 Conference

The conference test set deals with matching several ontologies on the same topic. It also features a consensus workshop aimed at studying the elaboration of consensus when establishing the reference alignments.

9.1 Test set

The Conference collection consists of fourteen ontologies (developed within the OntoFarm project[10]) in the domain of organizing conferences. In contrast to the last year's conference track, there are four new ontologies. 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 alignments 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.

[9] Silas was trained on a set of books which is different from the evaluation set we used.

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

– *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 provided either complete alignments or interesting correspondences (nuggets), for all or some pairs of ontologies. There was no reference alignment. Instead, organizers of this track offered manual a posteriori evaluation of results. Organizers also plan to offer a posteriori evaluation of results by data-mining techniques. Manual evaluation produced statistics such as precision and will serve as input into data-mining based evaluation. During manual evaluation some interesting correspondences were chosen as a background material for the consensus building discussion.

9.2 Results

During the evaluation phase, we manually labelled correspondences by several tags in order to enable further processing of the results. In part we used those tags for computing traditional precision and so-called *relative-recall* (see Figure 9), which is the ratio of the number of correct correspondences found by a system over the number of the correct correspondences found by any of the systems. Next we also counted two quite soft metrics: *ratioSubs* and *ratioTriv*, where ratioSubs shows ratio of the number of subsumption errors and the number of incorrect correspondences, and ratioTriv shows ratio of the number of the so-called trivial correspondences and the number of correct correspondences (see Figure 10). Trivial correspondences are correct correspondences where aligned concepts have the very same label, thus exact string matching can fully work. All results from this phase are available on the result report page[11]. Those global statistics more or less reflect the quality of results of participants. Moreover, these tags were suitable for choosing controversial correspondences as input to the Consensus Building Workshop where additional fine grain results were obtained (see next).

Participants differ in the number of alignments submitted for evaluation:

– 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.
– 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.

Consensus Building Workshop (CBW). As a part of the Ontology Matching workshop we organized for the second year the so-called Consensus Building Workshop. Motivation for this event is to thoroughly discuss controversial correspondences and collaboratively trying to achieve consensus about (in)correctness of correspondences.

[11] http://oaei.ontologymatching.org/2007/result/conference/

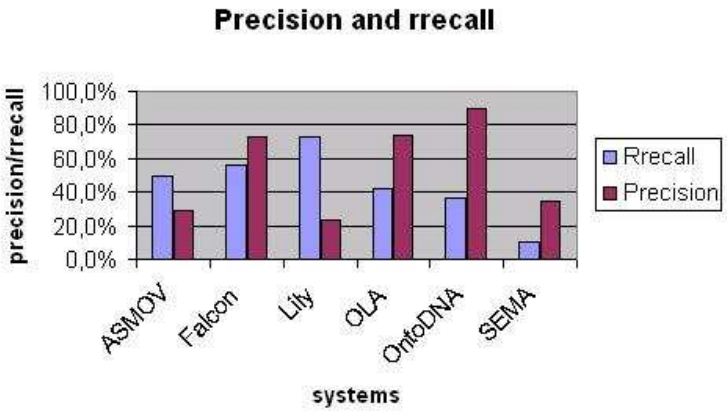

Fig. 9. Precision and relative-recall.

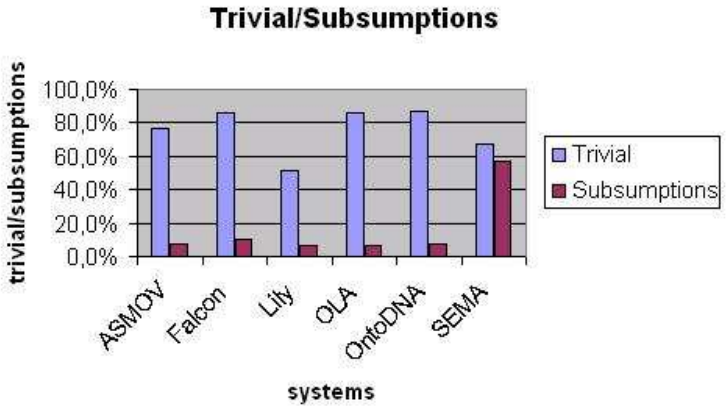

Fig. 10. ratioTriv and ratioSubs.

The controversial correspondences are those that were quite uncertain, unclear, interesting or obscure for different reasons during the process of evaluation. Intended results of this event are two-fold: (i) provide *feedback* for authors of the involved systems and (ii) examine the *argumentation process*. In comparison with 2006, participants of the discussion could argue for or against correspondences generally, and newly, we also considered application usage of correspondences. Organizers selected the application usage as a transformation application, i.e., final correspondences are to be used for conference data transformation from one software tool for organizing conference to another one. Altogether 15 correspondences were discussed. The board of experts finally achieved consensus for each of the correspondences. For some correspondences consensus was built easier, while for other argumentation had to be more sophisticated. The web site provides information about discussed controversial correspondences[12]. Additionally, for reaching agreement, visualization in the Protégé editor was used.

Some examples of the discussed correspondences include:

- `Individual_Presentation` and `Presentation`. Inspection of context, i.e., all subclasses, shows that `Individual_Presentation` is more general because `Presentation` contains just types of presentation format.
- `Name_of_conference` and `Topic_of_conference`. These concepts cannot be equivalent given common sense (or with regard to application-usage). On the other side, ontologies do not provide enough specification for distinguishing these concepts. In this case, it is needed to use some background knowledge.
- `has_author` and `submitted_by`. We concluded that these properties are equivalent according to their equivalent domains and ranges, i.e., how they are specified in the ontologies. However, regarding application-usage, they cannot be equivalent, because it can happen that sometimes the person who submitted the paper is not the author at the same time.
- `Presenter` and `Speaker`. In this case final consensus about equality of concepts was based on additional axiom: Paper presentedBy only Speaker.

Like in last CBW, the arguments are used in the same order: besides *lexical* reasons, the *context* of elements in question is considered primarily. Then, subclasses and superclasses which can unveil different extensions of classes are used. Finally, related properties and axioms are also considered.

CBW demonstrated that a board of experts can achieve agreement regarding (in)correctness of correspondences. CBW also shows that the problem can be with the meaning of concepts in terms of intensions of designers; some ontologies have not fully specified concepts. Application-usage of correspondences generally speeds up the discussion and makes the correspondence clearer.

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

[12] `http://oaei.ontologymatching.org/2007/result/conference/cbw07.ppt`

really taken into account that we identify with an asterisk (*). We reiterate those lessons that still apply with new ones:

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 improve these tests while preserving the comparability over years.
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) This year, we have detected (through random verifications) some submissions which were not strictly complying to the evaluation rules. This suggests to be more strict about control in future campaigns.
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.
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 evaluation of matching systems becomes more complex.

Most of the participants have provided description of their systems and their experience in the evaluation[13]. 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`.

[13] The SCARLET system is described in [12].

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 papers.

We are grateful to Henk Matthezing, Lourens van der Meij and Shenghui Wang who have made crucial contributions to implementation and reporting for the Library track. The evaluation at KB could not have been possible without the commitment of Yvonne van der Steen, Irene Wolters, Maarten van Schie, and Erik Oltmans.

The following persons were involved in the Food and Environment tasks: 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). In addition, Ondřej Šváb and Vojtěch Svátek were 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 17th 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. Antoine Isaac, Claus Zinn, Henk Matthezing, Lourens van der Meij, Stefan Schlobach, and Shenghui Wang. The value of usage scenarios for thesaurus alignment in cultural heritage context. In *Proceedings of the ISWC+ASWC International workshop on Cultural Heritage on the Semantic Web*, Busan (KR), 2007.
11. 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.
12. 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.
13. 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.
14. Willem Robert van Hage, Antoine Isaac, and Zharko Aleksovski. Sample evaluation of ontology-matching systems. In *Proceedings of the ISWC+ASWC International workshop on Evaluation of Ontologies and Ontology-based Tools*, Busan (KR), 2007.

Grenoble, Amsterdam, Trento, Mannheim, and Prague, December 10th, 2007