Results of the Ontology Alignment Evaluation Initiative 2008 [*]

Caterina Caracciolo[1], Jérôme Euzenat[2], Laura Hollink[3], Ryutaro Ichise[4], Antoine Isaac[3], Véronique Malaisé[3], Christian Meilicke[5], Juan Pane[6], Pavel Shvaiko[7], Heiner Stuckenschmidt[5], Ondřej Šváb-Zamazal[8], and Vojtěch Svátek[8]

[1] FAO, Roma, Italy
Caterina.Caracciolo@fao.org
[2] INRIA & LIG, Montbonnot, France
jerome.euzenat@inria.fr
[3] Vrije Universiteit Amsterdam, The Netherlands
{laurah,vmalaise,aisaac}@few.vu.nl
[4] National Institute of Informatics, Tokyo, Japan
ichise@nii.ac.jp
[5] University of Mannheim, Mannheim, Germany
{heiner,christian}@informatik.uni-mannheim.de
[6] University of Trento, Povo, Trento, Italy
pane@dit.unitn.it
[7] TasLab, Informatica Trentina, Trento, Italy
pavel.shvaiko@infotn.it
[8] 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-2008 builds over previous campaigns by having 4 tracks with 8 test sets followed by 13 participants. Following the trend of previous years, more participants reach the forefront. The 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 [7]. The main goal of the Ontology Alignment Evaluation Initiative is 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,

[*] This paper improves on the "First results" initially published in the on-site proceedings of the ISWC workshop on Ontology Matching (OM-2008). The only official results of the campaign, however, are on the OAEI web site.

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

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 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) [18]. 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], in 2006 at the first Ontology Matching workshop collocated with ISWC [6], and in 2007 at the second Ontology Matching workshop collocated with ISWC+ASWC [8]. Finally, in 2008, OAEI results were presented at the third Ontology Matching workshop collocated with ISWC, in Karlsruhe, Germany[2].

We have continued previous years' 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 2008 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 3-10 discuss in turn the settings and the results of each of the test cases. Section 11 overviews lessons learned from the campaign. Finally, Section 12 outlines future plans and Section 13 concludes.

2 General methodology

We first present the test cases proposed this year to OAEI participants. Then we describe the three steps of the OAEI campaign and report on the general execution of the campaign. In particular, we list participants and the tests they considered.

2.1 Tracks and test cases

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

The benchmark track (§3): Like in previous campaigns, a systematic benchmark series has been produced. The goal of this benchmark series is to identify the areas in which each matching algorithm is strong and 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 offers ontologies using OWL modeling capabiities:

- **Anatomy: (§4)** The anatomy real world case is about matching the Adult Mouse Anatomy (2744 classes) and the NCI Thesaurus (3304 classes) describing the human anatomy.

[2] http://om2008.ontologymatching.org

FAO (§5): The FAO test case is a real-life case aiming at matching OWL ontologies developed by the Food and Agriculture Organization of the United Nations (FAO) related to the fisheries domain.

The directories and thesauri track proposed web directories, thesauri and generally less expressive resources:

Directory (§6): The directory real world case consists of matching web sites directories (like open directory or Yahoo's). It is more than 4 thousand elementary tests.

Multilingual directories (§7): The mldirectory real world case consists of matching web site directories (such as Google, Lycos and Yahoo's) in different languages, e.g., English and Japanese. Data sets are excerpts of directories that contain approximately one thousand categories.

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.

Very large crosslingual resources (§9): This real world test case requires matching very large resources (vlcr) available on the web, viz. DBPedia, WordNet and the Dutch audiovisual archive (GTAA), DBPedia is multilingual and GTAA is in Dutch.

The conference track and consensus workshop (§10): 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 alignments as well as in interesting individual correspondences ("nuggets"), aggregated statistical observations and/or implicit design patterns. Organizers of this track offered diverse a priori and a posteriori evaluation of results. For a selected sample of correspondences, consensus was sought at the workshop and the process was tracked and recorded.

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

test	formalism	relations	confidence	modalities	language
benchmark	OWL	=	[0 1]	open	EN
anatomy	OWL	=	[0 1]	blind	EN
fao	OWL	=	1	expert	EN+ES+FR
directory	OWL	=	1	blind	EN
mldirectory	OWL	=	1	blind	EN+JP
library	SKOS, OWL	narrow-, exact-,	1	blind	EN+DU
vlcr	SKOS, OWL	broad-, relatedMatch	1	blind	EN+DU
conference	OWL-DL	=, ≤	[0 1]	blind+consensual	EN

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

Ontologies to be matched and (where applicable) alignments have been provided in advance during the period between May 19th and June 15th, 2008. 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 1st. The data sets did not evolve after this period.

2.3 Execution phase

During the execution phase, 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 [5]. 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 alignments provided by the participants and returned comparisons on these results.

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 1st. 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, for the first time, we had less participants than in the previous year (though still more than in 2006): 4 in 2004, 7 in 2005, 10 in 2006, 18 in 2007, and 13 in 2008. However, participants were able to enter nearly as many individual tasks as last year: 48 against 50.

We have had not enough time to systematically validate the results which had been provided by the participants, but we run a few systems and we scrutinized some of the results.

We summarize the list of participants in Table 2. Similar to previous years 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 is an even distribution of systems on tests (unlike last year when there were two groups of systems depending on the size of the ontologies). This years' participation seems to be weakly correlated with the fact that a test has been offered before.

Software	confidence	benchmark	anatomy	fao	directory	mldirectory	library	vlcr	conference
Anchor-Flood		√	√						
AROMA	√	√	√	√					
ASMOV	√	√	√	√	√				√
CIDER	√	√			√				
DSSim	√	√	√	√	√	√	√	√	√
GeRoMe		√							
Lily	√	√	√	√	√	√	√		√
MapPSO		√		√	√	√			
RiMOM	√	√	√	√	√	√			
SAMBO	√	√	√	√					
SAMBOdtf	√	√	√	√					
SPIDER	√	√							
TaxoMap		√	√		√		√		
Total=13		13	9	8	7	4	3	1	3

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 can still regret to have not enough time for performing tests and evaluations. This may explain why even participants with good results last year did not participate this year. The summary of the results track by track is provided in the following seven 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.

After remarks of last year we made two changes on the tests this year:

- tests #249 and 253 still had instances in the ontologies, these have been suppressed this year. Hence the test is more difficult than previous years;

- tests which scrambled all labels within the ontology (#201-202, 248-254 and 257-262), have been complemented by tests which respectively only scramble 20%, 40%, 60% and 80% of the labels. Globally, this makes the tests easier to solve.

The kind of expected alignments is still limited: they only match named classes and properties, they mostly use the "=" relation with confidence of 1. Full description of these tests can be found on the OAEI web site.

3.2 Results

All the 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.

Results in Table 3 show already that the three systems, which last year were leading, are still relatively ahead (ASMOV, Lily and RiMOM) with three close followers (AROMA, DSSim, and Anchor-Flood replacing Falcon, Prior+ and OLA$_2$ last year). 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.

This year again, 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). 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). They do not feature the curves of previous years since the test sets have been changed.

These results and those displayed in Figure 2 single out the same group of systems, ASMOV, Lily, and RiMOM which seem to perform these tests at the highest level of quality. So this confirms the leadership that we observed on raw results.

Like the two previous years, there is a gap between these systems and their followers. The gap between these systems and the next ones (AROMA, DSSim, and Anchor-Flood) has reformed. It was filled last year by Falcon, OLA$_2$, and Prior+ which did not participate this year.

We have also 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 two best systems on this basis are the same: ASMOV and Lily. Their results are very comparable but never identical to the results provided in the previous years by RiMOM (2006) and Falcon (2005).

system	refalign		edna		Aflood		AROMA		ASMOV		CIDER		DSSim		GeRoMe		Lily		MapPSO		RiMOM		SAMBO		SAMBOdtf		SPIDER		TaxoMap	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
2008																														
1xx	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.96	0.79	1.00	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	0.34
2xx	1.00	1.00	0.41	0.56	0.96	0.69	0.96	0.70	0.95	0.85	0.97	0.57	0.97	0.64	0.56	0.52	0.97	0.86	0.48	0.53	0.96	0.82	0.98	0.54	0.98	0.56	0.97	0.57	0.95	0.21
3xx	1.00	1.00	0.47	0.82	0.95	0.66	0.82	0.71	0.81	0.77	0.90	0.75	0.90	0.71	0.61	0.40	0.87	0.81	0.49	0.25	0.80	0.81	0.95	0.80	0.91	0.81	0.15	0.81	0.92	0.21
H-mean	1.00	1.00	0.43	0.59	0.97	0.71	0.95	0.70	0.95	0.86	0.97	0.62	0.97	0.67	0.60	0.58	0.97	0.88	0.51	0.54	0.96	0.84	0.99	0.58	0.98	0.59	0.81	0.63	0.91	0.22
Symmetric relaxed measures																														
H-mean	1.00	1.00	0.73	1.00	1.00	0.72	error		0.99	0.90	error		error		error		0.99	0.89	error		error		0.99	0.58	0.99	0.59	error		1.00	0.24
2007																														
1xx	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.96	0.79	1.00	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	0.34
2xx	1.00	1.00	0.41	0.56	0.96	0.69	0.96	0.70	0.95	0.85	0.97	0.57	0.97	0.64	0.56	0.52	0.97	0.86	0.48	0.53	0.96	0.82	0.98	0.54	0.98	0.56	0.97	0.57	0.95	0.21
3xx	1.00	1.00	0.47	0.82	0.95	0.66	0.82	0.71	0.81	0.77	0.90	0.75	0.90	0.71	0.61	0.40	0.87	0.81	0.49	0.25	0.80	0.81	0.95	0.80	0.91	0.81	0.15	0.81	0.92	0.21
H-mean	1.00	1.00	0.45	0.61	0.97	0.71	0.96	0.72	0.95	0.85	0.97	0.62	0.97	0.68	0.59	0.54	0.96	0.87	0.52	0.55	0.95	0.83	0.98	0.59	0.98	0.61	0.67	0.62	0.95	0.22

Table 3. Means of results obtained by participants on the benchmark test case (corresponding to harmonic means). The symmetric relaxed measure corresponds to the three relaxed precision and recall measure of [4]. The 2007 subtable corresponds to the results obtained on the results of 2007 tests only (suppressing the 20-40-60-80% alteration).

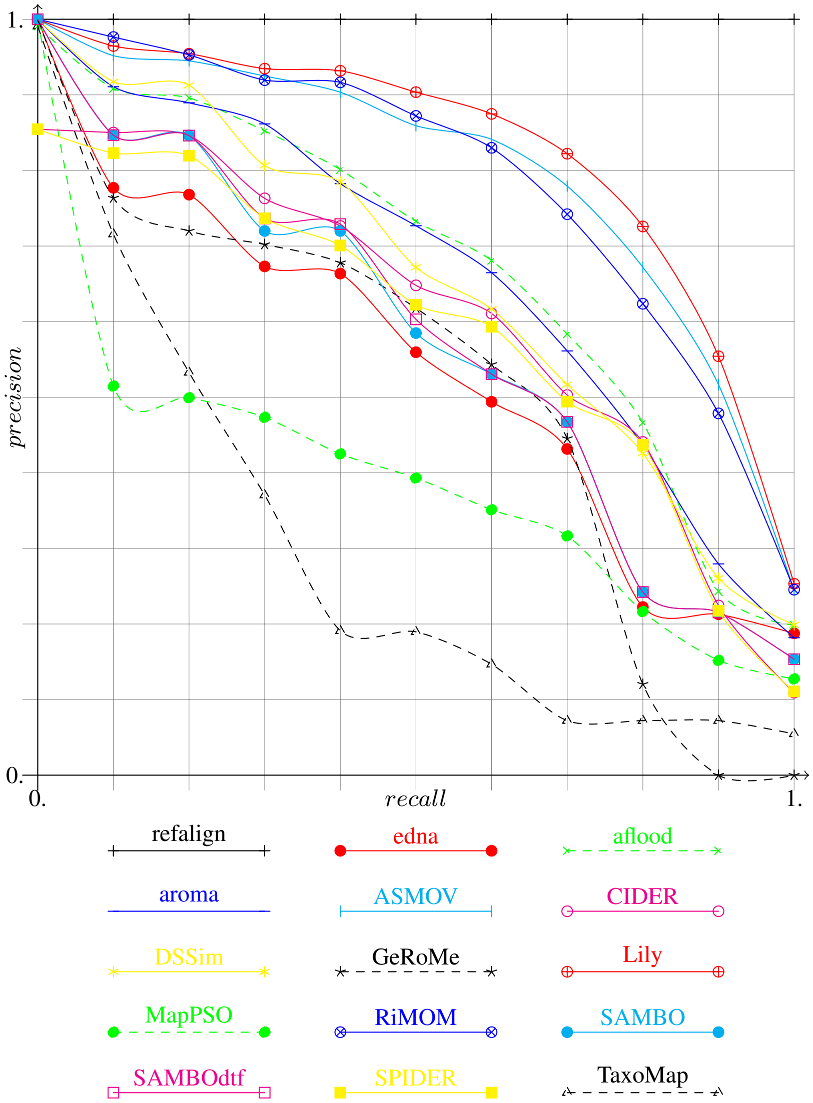

Fig. 1. Precision/recall graphs. They cut the results given by the participants under a threshold necessary for achieving $n\%$ recall and compute the corresponding precision. Systems for which these graphs are not meaningful (because they did not provide graded confidence values) are drawn in dashed lines. This is, as expected, those which have the lower results in these curves.

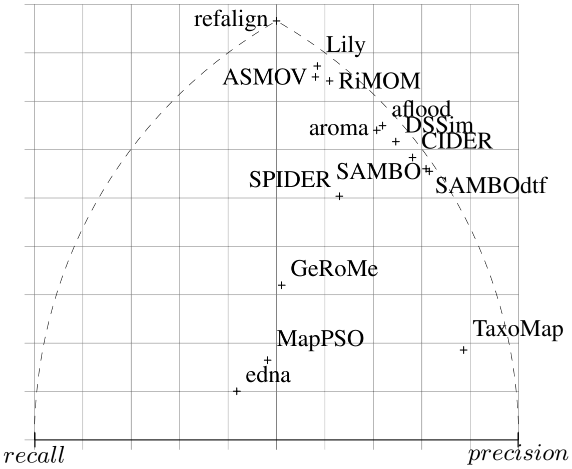

Fig. 2. Each point expresses the position of a system with regard to precision and recall.

Year	2004				2005		2006		2007				2008			
System	Fujitsu		PromptDiff		Falcon		RiMOM		ASMOV		Lily		ASMOV		Lily	
test	Prec.	Rec.	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	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	0.99	0.98	0.99	0.98
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	0.81	0.77	0.87	0.81
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	0.97	0.96	0.98	0.96

Table 4. Evolution of the best scores over the years on the basis of 2004 tests (RiMOM had very similar results to ASMOV's).

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.[3] Manually generating alignments between these ontologies requires an enormous effort by highly specialized domain experts. Supporting these experts by automatically providing correspondence proposals is challenging, due to the complexity and the specialized vocabulary of the domain.

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)[4], and the Adult Mouse Anatomical Dictionary[5], which has been developed as part of the Mouse Gene Expression Database project. Both resources are part of the Open Biomedical Ontologies (OBO). A more detailed description of the characteristics of the data set has already been given in the context of OAEI 2007 [8].

Due to the harmonization of the ontologies applied in the process of generating a reference alignment, a high number of rather trivial correspondences 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. The construction of the reference alignment has been described in [3]. To better understand the occurrence of non-trivial correspondences in alignment results, 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 98% precision and 61% recall. Notice that the value for recall is relatively high, which is partially caused by the harmonization process mentioned above. In 2007 we assumed that most matching systems would easily find the trivial correspondences. To our suprise this assumption has not been verified. Therefore, we applied again the additional measure referred to as $recall+$. $recall+$ 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 $recall+ = |(R \cap M) - S| \,/\, |R - S|$.

We divided the task of automatically generating an alignment into four subtasks. Task #1 is obligatory for participants of the anatomy track, while task #2, #3 and #4 are optional tasks. Compared to 2007 we also introduced #4 as challenging fourth subtask. For task #1 the 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. In particular,

[3] A large collection can be found at http://www.obofoundry.org/.

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

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

we are interested in how far matching systems improved their results compared to last years evaluation. For task #2 an alignment with increased precision has to be found. Contrary to this, in task #3 an alignment with increased recall has to be generated. We believe that systems configurable with respect to these requirements will be much more useful in concrete scenarios compared to static systems. While we expect most systems to solve the first three tasks, we expect only few systems to solve task #4. For this task a part of the reference alignment is available as additional input. In task #4 we tried to simulate the following scenario. Suppose that a group of domain experts already created an incomplete reference alignment by manually validating a set of automatically generated correspondences. As a result a partial reference alignment, in the following referred to as R_p, is available. Given both ontologies as well as R_p, a matching system should be able to exploit the additional information encoded in R_p. We constructed R_p as the union of the correct trivial correspondences and a small set of 54 non trivial correspondences. Thus R_p consists of 988 correspondences, while the complete reference alignment R contains 1523 correspondences.

4.2 Results

In total, nine systems participated in the anatomy task (in 2007 there were 11 participants). These systems can be divided into a group of systems using biomedical background knowledge and a group of systems that do not exploit domain specific background knowledge. SAMBO and ASMOV belong to the first group, while the other systems belong to the second group. Both SAMBO and ASMOV make use of UMLS, but differ in the way they exploit this additional knowledge. Table 5 gives an overview of participapating systems. In 2007 we observed that systems of the first group have a significant advantage of finding non trivial correspondences, in particular the best three systems (AOAS, SAMBO, and ASMOV) made use of background knowledge. We will later see whether this assumption could be verified with respect to 2008 submissions.

Compliance measures for task #1 Table 5 lists the results of the participants in descending order with respect to the achieved F-measure. In the first row we find the SAMBO system followed by its extension SAMBOdtf. SAMBO has achieved slightly better results for both precision and recall in 2008 compared to 2007. SAMBO now nearly reaches the F-measure 0.868 which AOAS achieved 2007. This is a notable result, since SAMBO is originally designed to generate alignment suggestions that are afterwards presented to a human evaluator in an interactive fashion. While SAMBO and SAMBOdtf make extensive use of biomedical background knowledge, the RiMOM matching system is mainly based on computing label edit-distances combined with similarity propagation strategies. Due to a major improvement of the RiMOM results, RiMOM is now one of the top matching systems for the anatomy track even though it does not make use of any specific background knowledge. Notice also that RiMOM solves the matching task in a very efficient way. Nearly all matching systems participating 2007 improved their results, while ASMOV and TaxoMap obtained slightly worse results. Further considerations have to clarify the reasons for this decline.

Task #2 and #3 As explained above these subtasks show in how far matching systems can be configured towards a trade-off between precision and recall. To our surprise only four participants submitted results for task #2 and #3 showing that they were able to

System	Runtime	BK	Precision	Recall	Recall+	F-Measure
SAMBO	≈ 12h	yes	0.869 0.845	0.836 0.797	0.586 0.601	0.852 0.821
SAMBOdtf	≈ 17h	yes	0.831	0.833	0.579	0.832
RiMOM	≈ 24min	no	0.929 0.377	0.735 0.668	0.350 0.404	0.821 0.482
aflood	1min 5s	no	0.874	0.682	0.275	0.766
Label Eq.	-	*no*	*0.981 0.981*	*0.613 0.613*	*0.000 0.000*	*0.755 0.755*
Lily	≈ 3h 20min	no	0.796 0.481	0.693 0.567	0.470 0.387	0.741 0.520
ASMOV	≈ 3h 50min	yes	0.787 0.802	0.652 0.711	0.246 0.280	0.713 0.754
AROMA	3min 50s	no	0.803	0.560	0.302	0.660
DSSim	≈ 17min	no	0.616 0.208	0.624 0.189	0.170 0.070	0.620 0.198
TaxoMap	≈ 25min	no	0.460 0.586	0.764 0.700	0.470 0.234	0.574 0.638

Table 5. Runtime, use of domain specific background knowledge (BK), precision, recall, recall+ and F-measure for task #1. Results of 2007 evaluation are presented in smaller font if available. Notice that the measurements of 2007 have been slightly corrected due to some minor modifications of the reference alignment.

adapt their system for different scenarios of application. These systems were RiMOM, Lily, ASMOV, and DSSim. A more detailed discussion of their results with respect to task #2 and #3 can be found on the OAEI anatomy track webpage[6].

System	Δ-Precision	Δ-Recall	Δ-F-Measure
SAMBO	$+0.024$ $_{0.636 \to 0.660}$	-0.002 $_{0.626 \to 0.624}$	$+0.011$ $_{0.631 \to 0.642}$
SAMBOdtf	$+0.040$ $_{0.563 \to 0.603}$	$+0.008$ $_{0.622 \to 0.630}$	$+0.025$ $_{0.591 \to 0.616}$
ASMOV	$+0.063$ $_{0.339 \to 0.402}$	-0.004 $_{0.258 \to 0.254}$	$+0.019$ $_{0.293 \to 0.312}$
RiMOM	$+0.012$ $_{0.700 \to 0.712}$	$+0.000$ $_{0.370 \to 0.370}$	$+0.003$ $_{0.484 \to 0.487}$

Table 6. Changes in precision, recall and F-measure based on comparing $M_1 \setminus R_p$ resp. $M_4 \setminus R_p$ with the unknown part of the reference alignment $R \setminus R_p$.

Task #4 Four systems participated in task #4. These systems were SAMBO and SAMBOdtf, RiMOM, and ASMOV. In the following we refer to an alignment generated for task #1 resp. #4 as M_1 resp. M_4. Notice first of all that a direct comparison between M_1 and M_4 is not appropriate to measure the improvement that results from exploiting R_p. We thus have to compare $M_1 \setminus R_p$ resp. $M_4 \setminus R_p$ with the unknown subset of the reference alignment $R_u = R \setminus R_p$. The differences between M_1 (partial reference alignment not available) and M_4 (partial reference alignment given) are presented in Table 6. All participants slightly increased the overall quality of the generated alignments with respect to the unknown part of the reference alignment. SAMBOdtf and ASMOV exploited the partial reference alignment in the most effective way. The measured im-

[6] http://webrum.uni-mannheim.de/math/lski/anatomy08/

provement seems to be only minor at first sight, but notice that all of the correspodences in R_u are non trivial due to our choice of the partial reference alignment. The improvement is primarily based on generating an alignment with increased precision. ASMOV for example increases its precision from 0.339 to 0.402. Only SAMBOdtf also profits from the partial reference alignment by a slightly increased recall. Obviously, the partial reference alignment is mainly used in the context of a strategy which filters out incorrect correspondences.

Runtime Even though the submitted alignments have been generated on different machines, we believe that the runtimes provided by participants are nevertheless useful and provide a basis for an approximate comparison. For the two fastest systems, namely aflood and AROMA, runtimes have been measured by the track organizers on the same machine (Pentium D 3.4GHz, 2GB RAM) additionally. Compared to last years competition we observe that systems with a high runtime managed to decrease the runtime of their system significantly, e.g. Lily and ASMOV. Amongst all systems AROMA and aflood, both participating for the first time, performed best with respect to runtime efficiency. In particular, the aflood system achieves results of high quality in a very efficient way.

4.3 Conclusions

In last years evaluation, we concluded that the use of domain related background knowledge is a crucial point in matching biomedical ontologies. This observation is supported by the claims made by other researchers [1; 15]. The current results partially support this claim, in particular the good results of the SAMBO system. Nevertheless, the results of RiMOM and Lily indicate that matching systems are able to detect non trivial correspondences even though they do not rely on background knowledge. To support this claim we computed the union of the alignments generated by RiMOM and Lily. As a result we measured that 61% of all non trivial correspondences are included in the resulting alignment. Thus, there seems to be a significant potential of exploiting knowledge encoded in the ontologies. A combination of both approaches might result in a hybrid matching strategy that uses both background knowledge and the internal knowledge to its full extent.

5 FAO

The Food and Agriculture Organization of the United Nations (FAO) collects large amounts of data about all areas related to food production and consumption, including statistical data, e.g., time series, and textual documents, e.g., scientific papers, white papers, project reports. For the effective storage and retrieval of these data sets, controlled vocabularies of various types (in particular, thesuri and metadata hierarchies) have extensively been used. Currently, this data is being converted into ontologies for the purpose of enabling connection between data sets otherwise isolated from one another. The FAO test case aims at exploring the possibilities of establishing alignments between some of the ontologies traditionally available. We chose a representative subset of them, that we describe below.

5.1 Test set

The FAO task involves the three following ontologies:

- AGROVOC[7] is a thesaurus about all matters of interest for FAO, it has been translated into an OWL ontology as a hierarchy of classes, where each class corresponds to an entry in the thesaurus. For technical reasons, each class is associated with an instance with the same name. Given the size and the coverage of AGROVOC, we selected only the branches of it that have some overlap with the other considered ontologies. We then selected the fragments of AGROVOC about "organisms," "vehicles" (including vessels), and "fishing gears".
- ASFA[8] is a thesaurus specifically dedicated to aquatic sciences and fisheries. In its OWL translation, descriptors and non-descriptors are modeled as classes, so the ontology does not contain any instance. The tree structure of ASFA is relatively flat, with most concepts not having subclasses, and a maximum depth of 4 levels. Concepts have associated annotations, each of which containing the English definition of the term.
- Two specific fisheries ontologies in OWL[9], that model coding systems for commodities and species, used as metadata for statistical time series. These ontologies have a fairly simple class structure, e.g., the species ontologies has one top class and four subclasses, but a large number of instances. They contain instances in up to 3 languages (English, French and Spanish).

Based on these ontologies, participats were asked to establish alignments between:

1. AGROVOC and ASFA (from now on called agrasfa),
2. AGROVOC and fisheries ontology about biological species (called agrobio),
3. the two ontologies about biological species and commodities (called fishbio).

Given the structure of the ontologies described above, the expectation about the resulting alignments was that the alignment between AGROVOC and ASFA (agrasfa) would be at the class level, since both model entries of the thesaurus as classes. Analogously, both the alignment between AGROVOC and biological species (agrobio), and the alignment between the two fisheries ontologies (fishbio) is expected to be at the instance level. However, no strict instructions were given to participants about the exact type of alignment expected, as one of the goals of the experiment was to find how automatic systems can deal with a real-life situation, when the ontologies given are designed according to different models and have little or no documentation.

The equivalence correspondences requested for the agrasfa and agrobio subtracks are plausible, given the similar nature of the two resources (thesauri used for human indexing, with some overlap in the domain covered). In the case of the fishbio subtrack this is not true, as the two ontologies involved are about two domains that are disjoint, although related, i.e., commodities and fish species. The relation between the two domains is that a specific species (or more than one) are the primary source of the goods

[7] http://www.fao.org/aims/ag_intro.htm

[8] http://www.fao.org/fishery/asfa/8

[9] http://www.fao.org/aims/neon.jsp

sold, i.e. the commodity. Their relation then is not an equivalence relation but can rather be seen, in OWL terminology, as an object property with domain and range sitting in different ontologies. The intent of the subtrack fishbio is then to explore the possibility of using the machinery available for inferring equivalence correspondence to non conventional cases.

5.2 Evaluation procedure

All participants but one, Aroma, returned equivalence correspondence only. The non-equivalence correspondences of Aroma were ignored.

A reference alignment was obtained by randomly selecting a specific number of correspondences from each system and then pooling together. This provided a sample alignment A^0.

This sample alignment was evaluated by FAO experts for correctness. This provided a partial reference alignment R^0. We had two assessors: one specialized in thesauri and daily working with AGROVOC (assessing the alignments of the track agrasfa) and one specialized in fisheries data (assessing subtracks agrobio and fishbio). Given the differences between the ontologies, some transformations had to be made in order to present data to the assessors in a user-friendly manner. For example, in the case of AGROVOC, evaluators were given the English labels together with all available "used for" terms (according to the thesauri terminology familiar to the assessor).

dataset	retrieved (A^*)	evaluated (A^0)	correct (R^0)	(A^0/A^*)	(R^0/A^0)
agrasfa	2588	506	226	.19	.45
agrobio	742	264	156	.36	.59
fishbio	1013	346	131	.26	.38
TOTAL	4343	1116	513	.26	.46

Table 7. Size of returned results and samples.

Table 7 summarizes the sample size per each data sets. The second column (retrieved) contains the total number of distinct correspondences provided by all participants for each track. The third column (evaluated) reports the size of the sample extracted for manual assessment. The forth column (correct) reports the number of correspondences found correct by the assessors.

After manual evaluation, we realized that some participants did not use the correct URI in the agrasfa dataset, so some correspondences were considered as different even though they were actually the same. However, this happened only in very few cases.

For each system, precision was computed on the basis of the subset of alignments that were manually assessed, i.e., $A \cap A^0$. Hence,

$$P^0(A, R^0) = P(A \cap A^0, R^0) = |A \cap R^0|/|A \cap A^0|$$

The same was considered for recall which was computed with respect to the total number of correct correspondences per subtrack, as assessed by the human assessors. Hence,

$$R^0(A, R^0) = R(A \cap A^0, R^0) = |A \cap R^0|/|R^0|$$

Recall is expected to be higher than actual recall because it is based only on correspondences that at least one system returned, leaving aside those that no system were able to return.

We call these two measures relative precision and recall because they are relative to the sample that has been extracted.

5.3 Results

Table 8 summarizes the precision and (relative) recall values of all systems, by subtracks. The third column reports the total number of correspondences returned by each system per subtrack. All non-equivalence correspondences were discarded, but this only happened for one systems (Aroma). The fourth column reports the number of alignments from the system that were evaluated, while the fifth column reports the number of correct alignments as judged by the assessors. Finally, the sixth and seventh columns report the values of relative precision and recall computed as described above.

		retrieved	evaluated	correct	RPrecision	RRecall
System	subtrack	$\|A\|$	$\|A \cap A^0\|$	$\|A \cap R^0\|$	$P^0(A, R^0)$	$R^0(A, R^0)$
Aroma	agrasfa	195	144	90	0.62	0.40
	agrobio	2	4	0		
	fishbio	11				
ASMOV	agrafsa	1				
	agrobio	0				
	fishbio	5				
DSSim	agrasfa	218	129	70	0.54	0.31
	agrobio	339	214	151	0.71	0.97
	fishbio	243	166	79	0.48	0.60
Lily	agrasfa	390	105	91	0.87	0.40
MapPSO	agrobio*	6				
	fishbio*	16				
RiMOM	agrasfa	743	194	158	0.81	0.70
	agrobio	395	219	149	0.68	0.95
	fishbio	738	217	118	0.54	0.90
SAMBO	agrasfa	389	176	121	0.69	0.53
SAMBOdtf	agrasfa	650	219	124	0.57	0.55

Table 8. Participant results per datasets. The star (*) next to a system marks those systems which matched properties.

One system (MapPSO) returned alignments of properties, which were discarded and therefore no evaluation is provided in the table. The results of ASMOV were also not evaluated because too few to be considered. Finally, the evaluation of Aroma is incomplete due to the non equivalence correspondence returned, that were discarded before pooling the results together to create the reference alingment.

5.4 Discussion

The sampling method that has been used is certainly not perfect. In particular, it did not allow to evaluate two systems which returned few results (ASMOV and MapPSO). However, the results returned by these system were not likely to provide good recall.

Moreover, the very concise instructions and the particular character of the test sets, clearly puzzled participants and their systems. As a consequence, the results may not be as good as if the systems were applied to polished tests with easily comparable data sets. This provides a honest insight of what these systems would do when confronted with these ontologies on the web. In that respects, the results are not bad.

From DSSim and RiMOM results, it seems that fishbio is the most difficult task in terms of precision and agrasfa the most difficult in terms of recall (for most of the systems). The fact that only two systems returned usable results for agrobio and fishbio makes comparison of systems very difficult at this stage. However, it seems that RiMOM is the one that provided the best results. RiMOM is especially interesting in this real-life case, as it performed well both when an alignment between classes and an alignment between instances is appropriate. Given the fact that in real-life situations it is rather common to have ontologies with a relatively simple class structure and a very large population of instances, this is encouraging.

6 Directory

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

6.1 Test set

The data set exploited in the directory matching task was constructed from Google, Yahoo and Looksmart web directories following the methodology described in [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 [9], 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.

6.2 Results

In OAEI-2008, 7 out of 13 matching systems participated on the web directories test case, while in OAEI-2007, 9 out of 18, in OAEI-2006, 7 out of 10, and in OAEI-2005, 7 out of 7 did it.

Precision, recall and F-measure results of the systems are shown in Figure 3. These indicators have been computed following the TaxMe2 [9] methodology, with the help of Alignment API [5], version 3.4.

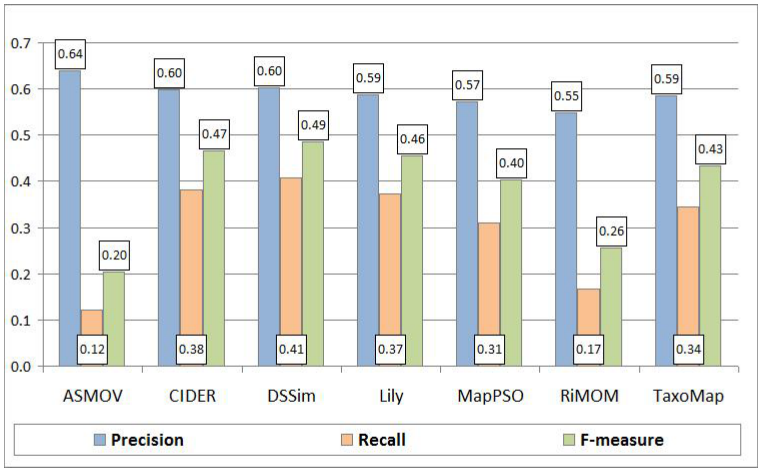

Fig. 3. Matching quality results.

We can observe from Table 9, that all the systems that participated in the directory track in 2007 and 2008 (ASMOV, DSSim, Lily and RiMOM), have increased their precision values. Considering recall, we can see that in general the systems that had participated in 2007 and 2008 directory tracks, have decreased their values, the only system that increased its recall values is DSSim. In fact, DSSim is the system with the highest F-measure value in 2008.

Table 9 shows that in total 21 matching systems have participated during the 4 years (2005 - 2008) of the OAEI campaign in the directory track. No single system has participated in all campaigns involving the web directory dataset (2005 - 2008). A total of 14 systems have participated only one time in the evaluation, 5 systems have participated 2 times, and only 2 systems have participated 3 times. The systems that have participated in 3 evaluations are Falcon (2005, 2006 and 2007) and RiMoM (2006, 2007, 2008), the former with a constant increase in the quality of the results, the later with a constant increase in precision, but in the last evaluation (2008) recall dropped significantly from 71% in 2007, to 17% in 2008.

System	**Recall**				**Precision**			**F-Measure**		
Year →	2005	2006	2007	2008	2006	2007	2008	2006	2007	2008
ASMOV			0.44	0.12		0.59	0.64		0.50	0.20
automs		0.15			0.31			0.20		
CIDER				0.38			0.60			0.47
CMS	0.14									
COMA		0.27			0.31			0.29		
ctxMatch2	0.09									
DSSim			0.31	0.41		0.60	0.60		0.41	0.49
Dublin20	0.27									
Falcon	0.31	0.45	0.61		0.41	0.55		0.43	0.58	
FOAM	0.12									
hmatch		0.13			0.32			0.19		
Lily			0.54	0.37		0.57	0.59		0.55	0.46
MapPSO				0.31			0.57			0.40
OCM		0.16			0.33			0.21		
OLA	0.32		0.84			0.62			0.71	
OMAP	0.31									
OntoDNA			0.03			0.55			0.05	
Prior		0.24	0.71		0.34	0.56		0.28	0.63	
RiMOM		0.40	0.71	0.17	0.39	0.44	0.55	0.40	0.55	0.26
TaxoMap				0.34			0.59			0.43
X-SOM			0.29			0.62			0.39	
Average	*0.22*	*0.26*	*0.50*	*0.30*	*0.35*	*0.57*	*0.59*	*0.29*	*0.49*	*0.39*
#	*7*	*7*	*9*	*7*	*7*	*9*	*7*	*7*	*9*	*7*

Table 9. Summary of submissions by year (no precision was computed in 2005). The Prior line covers Prior+ as well and the OLA line covers OLA_2 as well.

As can be seen in Figure 4 and Table 9, there is an increase in the average precision for the directory track of 2008, along with a decrease in the average recall compared to 2007. Notice that in 2005 the data set allowed only the estimation of recall, therefore Figure 4 and Table 9 do not contain values of precision and F-measure for 2005.

A comparison of the results in 2006, 2007 and 2008 for the top-3 systems of each year based on the highest values of the F-measure indicator is shown in Figure 5. The key observation here is that unfortunately the top-3 systems of 2007 did not participate in the directory task this year, therefore, the top-3 systems for 2008 is a new set of systems (Lily, CIDER and DSSim). From these 3 systems, CIDER is a newcomer, but Lily and DSSim had also participated in the directory track of 2007, when they did not manage to enter into the top-3 list.

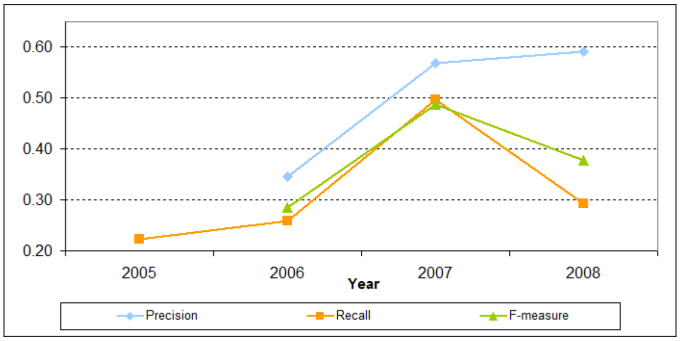

Fig. 4. Average results of the top-3 systems per year.

The quality of the best F-measure result of 2008 (0.49) demonstrated by DSSim is lower than the best F-measure of 2007 (0.71) by OLA$_2$ but still higher than that of 2006 by Falcon (0.43). The best precision result of 2008 (0.64) demonstrated by ASMOV is higher than the results obtained in 2007 (0.62) by both OLA$_2$ and X-SOM. Finally, for what concerns recall, the best result of 2008 (0.41) demonstrated by DSSim is also lower than the best results obtained in 2007 (0.84) obtained by OLA$_2$ and in 2006 (0.45) by Falcon. This decrease in the maximum values achieved by the participating systems may be caused by participants tuning their system parameters for more diverse tasks this year. Hence, the overall results of systems could have improved at the expense of results in the directory track. For example, we can observe that both ASMOV and Lily have very good results (over 90% F-measure) for the Benchmark-2008 track, which are higher than the Benchmarck-2007 track.

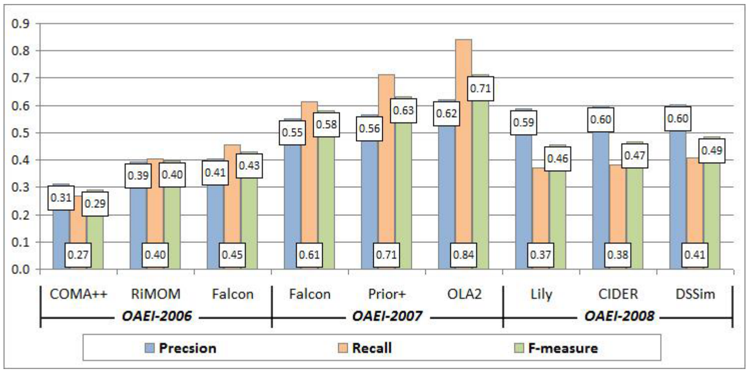

Fig. 5. Comparison of matching quality results in 2006, 2007 and 2008.

Partitions of positive and negative correspondences according to the system results are presented in Figure 6 and Figure 7, respectively.

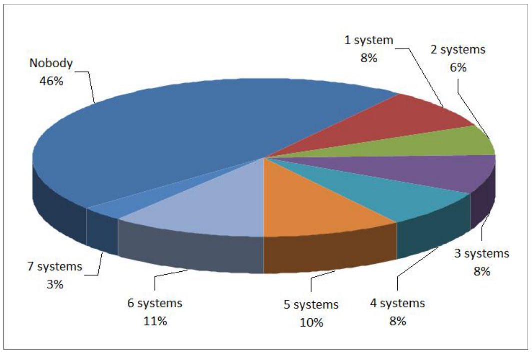

Fig. 6. Partition of the system results on positive correspondences.

Figure 6 shows that the systems managed to discover only 54% of the total number of positive correspondences (Nobody = 46%). Only 11% of positive correspondences were found by almost all (6) matching systems, while 3% of the correspondences were found by all the participants in 2008. This high percentage of positive correspondences not found by the systems correspond to the low recall values we observe in Table 9, which are the cause of the decrease in average recall from 2007 to 2008. Figure 7 shows that most of the negatives correspondences were not found by the systems (correctly). Figure 7 also shows that six systems found 11% of negative correspondences, i.e., mistakenly returned them as positive. The last two observations suggest that the discrimination ability of the dataset remains still high as in previous years.

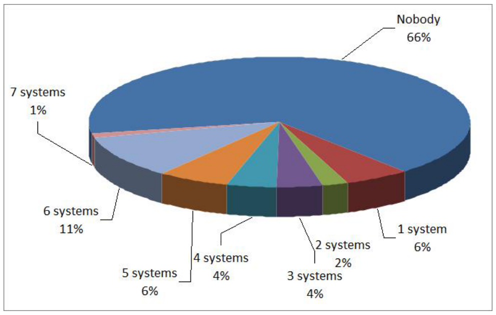

Fig. 7. Partition of the system results on negative correspondences.

Let us now compare partitions of the system results in 2006, 2007 and 2008 on positive and negative correspondences, see Figure 8 and Figure 9, respectively.

Figure 8 shows that 46% of 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. In 2008, 46% of the positive correspondences have not been found by the participating systems, as in 2006. This year, systems performed in the line of 2006. In 2007, the results were exceptional because the participating systems alltogether had a full coverage of the expected results and very high precision and recall. Unfortunately, the best systems of last year did not participate this year and the other systems do not seem to cope with the previous results.

Figure9 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 2008 the value increased to 66%, this is, the set of participating systems in 2008 cor-

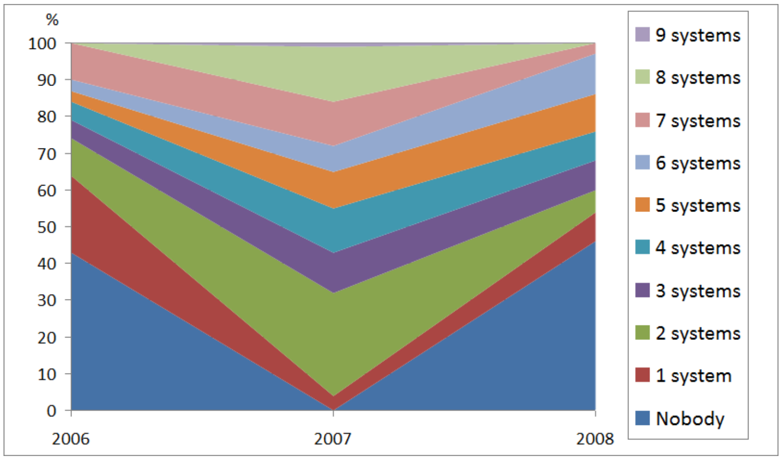

Fig. 8. Comparison of partitions of the system results on positive correspondences in 2006, 2007 and 2008.

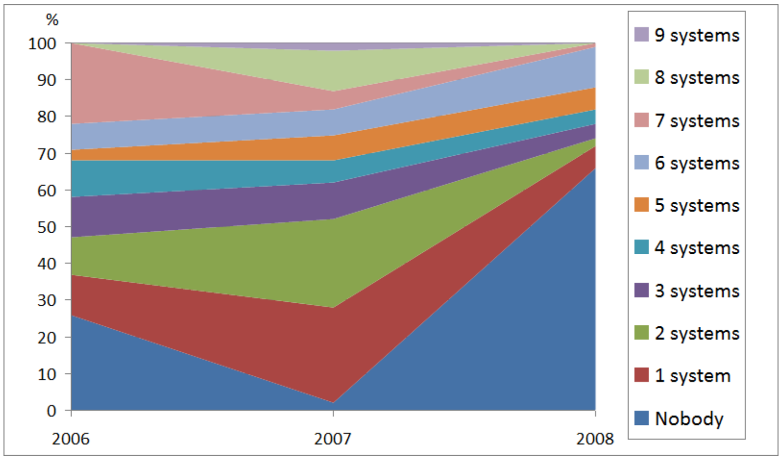

Fig. 9. Comparison of partitions of the system results on negative correspondences in 2006, 2007 and 2008.

rectly avoid more negative correspondences than those participating in 2006 and 2007. In 2006, 22% of negative correspondences were mistakenly found by all (7) the matching systems, while in 2007, this indicator decreased to 5% (for 7 systems), and in 2008, the value decreased even more to 1%. An interpretation of these observations could be that the set of participating systems in 2008 have a more "cautious" strategy than in 2007 and 2006. In 2007 we can observe that the set systems showed a more "brave" strategy in discovering correspondences, were the set of positive correspondences was fully covered, but covering mistakenly also 98% of the negative correspondences, while in 2008 the set of participating systems covered just 54% of the positive correspondences, but covering only 34% of negative correspondences.

6.3 Comments

An important observation from this evaluation is that ontology matching is still making progress on the web directory track this year, if we consider that the set of participating systems in 2008 is almost completely different compared to 2007. With respect to the average performance of the systems (given by F-Measure in Figure 4), the set of participating systems in 2008 performed worse than the set of participating systems in 2007, but better than those participating in 2006. This suggests that the systems participating in 2008 experienced a higher number of difficulties on the test case, in comparison to 2007, which means that there is still room for further improvements, specially in recall. A considerable remark this year is that it is hard for a single system to perform well in all the situations when finding correspondences is needed (which are simulated by the different OAEI tracks); this suggests that a general purpose matching system is difficult to construct. Finally, as partitions of positive and negative correspondences indicate (see Figure 6 and Figure 7), the dataset still retains a good discrimination ability, i.e., different sets of correspondences are still hard for the different systems.

7 Multilingual directories

The multilingual directory data set (mldirectory) is a data set created from real internet directory data. This data provides alignment problems for different internet directories. This track mainly fpcuses on multilingual data (English and Japanese) and instances.

7.1 Test data and experimental settings

The multilingual directory data set is constructed from Google (open directory project), Yahoo!, Lycos Japan, and Yahoo! Japan. The data set consists of five domains: automobile, movie, outdoor, photo and software, which are used in [11; 10]. There are four files for each domain. Two are for English directories and the rest are for Japanese directories. Each file is written in OWL. A file is organized into two parts. The first part describes the class structures, which are organized with `rdfs:subClassOf` relationships. Each class might also have `rdfs:seeAlso` properties, which indicate related classes. The second part is the description of instances of the classes. Each description has an instance ID, class name, instance label, and short description.

There are two main differences between the *mldirectory* data set and *directory* data set, which is also available for OAEI-2008.

- The first one is a multilingual set of directory data. As we mentioned above, the data set has four different ontologies with two different languages for one domain. As a result, we have six alignment problems for one domain. These include one English-English alignment, four English-Japanese alignments, and one Japanese-Japanese alignment.
- The second difference is the instances of classes. In the multilingual directory data set, the data not only has relationships between classes but also instances in the classes. As a result, we can use snippets of web pages in the Internet directories as well as category names in the multilingual directory data set.

We encouraged participants to submit alignments for all domains. Since there are five domains and each domain has six alignment patterns, this is thirty alignments in total. However, participants can submit some of them, such as the English-English alignment only.

Participants are allowed to use background knowledge such as Japanese-English dictionaries and WordNet. In addition, participants can use different data included in the multilingual directory data set for parameter tuning. For example, the participants can use automobile data for adjusting the participant's system, and then induce the alignment results for movie data by the system. Participants cannot use the same data to adjust their system, because the system will consequently not be applicable to unseen data. In the same manner, participants cannot use specifically crafted background knowledge because it will violate the assumption that we have no advanced knowledge of the unseen data.

7.2 Results

In the 2008 campaign, four participants dealt with the mldirectory data set: DSSim, Lily, MapPSO and RiMOM. Among the four systems, three of them – DSSim, MapPSO, and RiMOM – were used for all five domains in the English-English alignment, and one of them, Lily, was used in the task for two domains, automobile and movie. The number of correspondences found by the systems are shown in Table 10. As can be seen in this table, Lily finds more correspondences than do the other systems. Conversely, MapPSO retrieves only a few correspondences from the data set.

In order to learn the different biases of the systems, we counted the number of common correspondences retrieved by the systems. The results are shown in Table 11. The letters D, L, M and R in the top row denote system names DSSim, Lily, MapPSO, and RiMOM, respectively. For example, the DR column is the number of correspondences retrieved by both DSSim and RiMOM. We can see that both systems retrieve the same 82 correspondences in the movie domain. In this table, we see interesting phenomena. Lily and RiMOM have the same bias. For example, in the auto domain, 33% of the correspondences found by Lily were also retrieved by RiMOM, and 46% of the correspondences found by RiMOM were also retrieved by Lily. The same phenomenon is

	DSSim	Lily	MapPSO	RiMOM
Auto	188	377	265	275
Movie	1181	1864	183	1681
Outdoor	268	-	10	538
Photo	141	-	38	166
Software	372	-	60	536
Total	2150	2241	556	3196

Table 10. Number of correspondences found (English-English alignments).

also seen in the movie domain. In contrast, MapPSO has a very different tendency. Although the system found 556 alignments in total, only one correspondence was found by the other systems.

	D	L	M	R	DL	DM	DR	LM	LR	MR	DLM	DLR	DMR	LMR	DLMR
Auto	139	208	264	104	5	0	7	0	126	0	0	37	1	0	0
Movie	946	988	183	734	11	0	82	0	723	0	0	142	0	0	0
Outdoor	260	0	10	530	0	0	8	0	0	0	0	0	0	0	0
Photo	137	0	38	162	0	0	4	0	0	0	0	0	0	0	0
Software	338	0	60	502	0	0	34	0	0	0	0	0	0	0	0

Table 11. Number of common correspondences retrieved by the systems. D, L, M, and R denote DSSim, Lily, MapPSO, and RiMOM, respectively.

We also created a component bar chart (Figure 10) for clarifying the sharing of retrieved correspondences. In the automobile and movie domains, 80% of the correspondences are found by only one system, and most of the other 20% are found by both Lily and RiMOM. From this graph, we can see that Lily has the same bias as RiMOM, but the system still found many correspondences that the other systems did not find. For the remaining domains, outdoor, photo and software, the correspondences found by only one system reached almost 100%.

Unfortunately, the results of other alignment tasks such as English-Japanese alignments (ontology 1-3, ontology 1-4, ontology 2-3, and ontology 2-4), Japanese-Japanese alignments (ontology 3-4) were only submitted by RiMOM. The number of alignments by RiMOM are shown in Table 12.

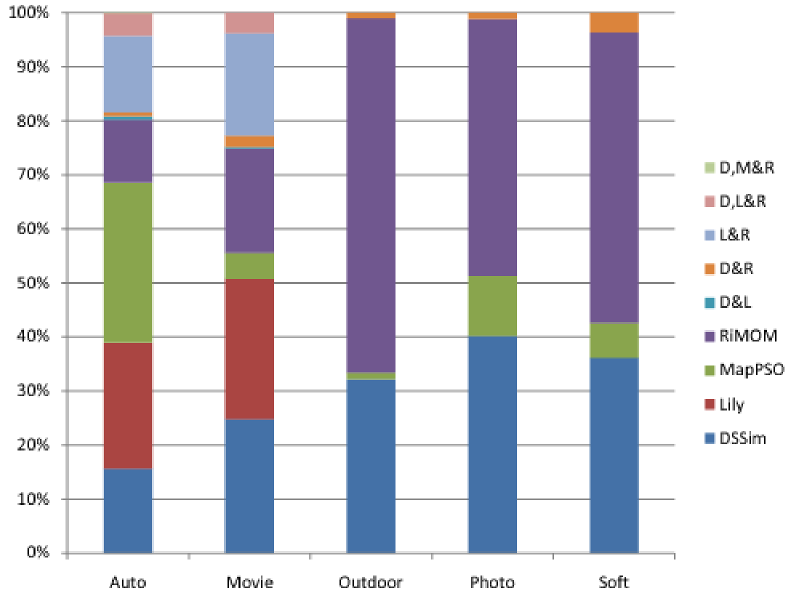

Fig. 10. Shared correspondences.

Domain	ont 1-2	ont 1-3	ont 1-4	ont 2-3	ont 2-4	ont 3-4	Total
Auto	275	99	242	79	225	262	1182
Movie	1681	35	30	35	59	65	1905
Outdoor	538	25	64	25	97	31	780
Photo	166	15	17	15	31	20	264
Software	536	104	125	78	100	84	1027

Table 12. Number of alignments by RiMOM.

8 Library

8.1 Data set

This test case deals with two large Dutch thesauri. The National Library of the Netherlands (KB) maintains two large collections of books: the Scientific Collection and the Deposit collection, containing respectively 1.4 and 1 million books. Each collection is annotated – *indexed* – using its own controlled vocabulary. The former is described using the GTT thesaurus, a huge vocabulary containing 35,194 general concepts, ranging from "Wolkenkrabbers" (Sky-scrapers) to "Verzorging" (Care). The latter is 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.

Each concept has exactly one preferred label, plus synonyms, extra hidden labels or scope notes. The language of both thesauri is Dutch,[10] 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. The thesauri's structural information is thus very poor.

For the purpose of the OAEI campaign, the two thesauri were made available in SKOS format. OWL versions were also provided, according to the – lossy – conversion rules detailed on the web site[11].

In addition, we have provided participants with *book descriptions*. At KB, almost 250000 books belong both to KB Scientific and Deposit collections, and are therefore already indexed against both GTT and Brinkman. Last year, we have used these books as a reference for evaluation. However, these books can also be a precious hint for obtaining correspondences. Indeed one of last year's participant had exploited co-occurrence of concepts, though on a collection obtained from another library. This year, we split the 250000 books in two sets: two third of them are provided to participants for alignment computation, and one third is kept as a test set to be used as a reference for evaluation.

8.2 Evaluation and results

Three systems provided final results: DSSim (2,930 `exactMatch` correspondences), Lily (2,797 `exactMatch` correspondences) and TaxoMap (1,872 `exactMatch` correspondences, 274 `broadMatch`, 1,031 `narrowMatch` and 40 `relatedMatch` correspondences).

We have followed the scenario-oriented approach followed for 2007 library track, as explained in [12].

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

[10] A quite substantial part of GTT concepts (around 60%) also have English labels.

[11] `http://oaei.ontologymatching.org/2008/skos2owl.html`

thesauri. Evaluation in such a context requires assessing the validity of each individual correspondence, as in "standard" alignment evaluation.

As last year, there was no reference alignment available. 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 – this is similar to the pooling approach that is used in major Information Retrieval evaluations, like TREC.

For manual evaluation, the set of all *equivalence* correspondences[12] was partitioned into parts unique to each combination of participant alignments, and each part was sampled. A total of 403 correspondences were assessed by one Dutch native expert.

From these assessments, precision and pooled recall were calculated with their 95% confidence intervals, taking into account sampling size. The results are shown in Table 13, which identifies DSSim as performing better than both other participants.

Alignment	Precision	Pooled recall
DSSim	93.3% ± 0.3%	68.0% ± 1.6%
Lily	52.9% ± 3.0%	36.8% ± 2.2%
TaxoMap (exactMatch)	88.1% ± 0.8%	41.1% ± 1.0%

Table 13. Precision and coverage for the thesaurus merging scenario.

DSSim has performed better than last year. This result stems probably from DSSim now proposing almost only exact lexical matches of SKOS labels, as opposed to last year.

For the sake of completeness, we also evaluated the precision of the TaxoMap correspondences that are not of type `exactMatch`. We categorized them according to the strength that TaxoMap gave them (0.5 or 1). 20% (±11%) of the correspondences with strength 1 are correct. The figure rises to 25.1% (±8.3%) when considering all non-`exactMatch` correspondences, which hints at the strength not being very informative.

Evaluation in an annotation translation scenario. The second usage scenario is based on an *annotation translation* process supporting the re-indexing of GTT-indexed books with Brinkman concepts [12].

This evaluation scenario interprets the correspondences provided by the different participants as rules to translate existing GTT book annotations into equivalent Brinkman annotations. Based on the quality of the results for books we know the correct annotations of, we can assess the quality of the initial correspondences.

[12] We did not proceed with manual evaluation of the *broader*, *narrower* and *related* links at once, as only one contestant provided such links.

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 2,930 rules for DSSim, 2,797 rules for Lily and 1,851 rules for TaxoMap. In addition, TaxoMap produces resp. 229, 897 and 39 rules considering `broadMatch`, `narrowMatch` and `relatedMatch`.

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.

Manual evaluation. Last year, we evaluated the results of the participants in two ways, one manual – KB indexers evaluating the generated indices – and one automatic – using books indexed against both GTT and Brinkman. This year, we have not performed manual investigation. Findings of last year can be found in [12].

Automatic evaluation and results. Here, the reference set consists of 81,632 dually-indexed books forming the test set presented in Section 8.1. The existing Brinkman indices from these books are taken as a reference to which the results of annotation translation are automatically compared.

The upper part of Table 14 gives an overview of the evaluation results when we only use the `exactMatch` correspondences. DSSim and TaxoMap perform similarly in precision, and much ahead of Lily. If precision almost reaches last year's best results, recall is much lower. Less than one third of the books were given at least one correct Brinkman concept in the DSSim case. At the annotation level, half of the translated concepts are not validated, and more than 75% of the real Brinkman annotation is not found. We already pointed out that the correspondences from DSSim are mostly generated by lexical similarity. This indicates, as last year, that lexically equivalent correspondences alone do not solve the annotation translation problem.

Participant	P_b	R_b	P_a	R_a	J_a
DSSim	56.55%	31.55%	48.73%	22.46%	19.98%
Lily	43.52%	15.55%	39.66%	10.71%	9.97%
TaxoMap	52.62%	19.78%	47.36%	13.83%	12.73%
TaxoMap+broadMatch	46.68%	19.81%	40.90%	13.84%	12.52%
TaxoMap+hierarchical	45.57%	20.23%	39.51%	14.12%	12.67%
TaxoMap+all correspondences	45.51%	20.24%	39.45%	14.13%	12.67%

Table 14. Results of annotation translations generated from correspondences.

Among the three participants, only TaxoMap generated `broadMatch` and `narrowMatch` correspondences. To evaluate their usefulness for annotation translation, we evaluated their influence when they were added to a common set of rules. As shown in the four *TaxoMap* lines in Table 14, the use of `broadMatch`, `narrowMatch` and `relatedMatch` correspondences slightly increases the chances of having a book given a correct annotation. However, this unsurprisingly results in a loss of precision.

8.3 Discussion

The first comment on this track concerns the *form* of the alignment returned by the participants, especially with respect to the type and cardinality of alignments. All three participants proposed alignments using the SKOS links we asked for. However, only one participants proposed hierarchical `broader`, `narrower` and `related` links. Experiments show that these links can be 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.

This actually mirrors what happened in last year's campaign, where only one participant had given non-exact correspondence links – even though it was `relatedMatch` then. Evaluation had shown that even though the general quality was lowered by considering them, the loss of precision was not too important, which could make these links interesting for some application variants, *e.g.* semi-automatic re-indexing.

Second, there is no precise handling of one-to-many or many-to-many alignments, as last year. 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.

Finally, one shall notice the low coverage of alignments with respect to the thesauri, especially GTT: in the best case, only 2,930 of its 35K concepts were linked to some Brinkman concept, which is less than last year (9,500). This track, arguably because of its Dutch language context, is difficult. We had hoped that the release of a part of the set of KB's dually indexed books would help tackle this difficulty, as previous year's campaign had shown promising results when exploiting real book annotations. Unfortunately none of this year's participants have used this resource.

9 Very large crosslingual resources

The goal of the Very Large Crosslingual Resources task is twofold. First, we are interested in the alignment of vocabularies in different languages. Many collections throughout Europe are indexed with vocabularies in languages other than English. These collections would benefit from an alignment to resources in other languages to broaden the user group, and possibly enable integrated access to the different collections.

Second, we intend to present a realistic use case in the sense that the resources are large, rich in semantics but weak in formal structure, i.e., realistic on the Web. For collections indexed with an in-house vocabulary, the link to a widely-used and rich resource can enhance the structure and increase the scope of the in-house thesaurus.

9.1 Data set

Three resources are used in this task:

GTAA The GTAA is a Dutch thesaurus used by the Netherlands Institute for Sound and Vision to index their collection of TV programs. It is a facetted thesaurus, of which we use the following four themes: (1) **Subject**: the topic of a TV program, $\approx$ 3800 terms; (2) **People**: the main people mentioned in a TV program, $\approx$ 97.000 terms; **Names**: the main "Named Entities" mentioned in a TV program (Corporation names, music bands, etc.), $\approx$ 27.000 terms; **Location**: the main locations mentioned in a TV program or the place where it has been created, $\approx$ 14.000 terms.

WordNet WordNet is a lexical database of the English language developed at Princeton University[13]. Its main building blocks are synsets: groups of words with a synonymous meaning. In this task, the goal is to match noun-synsets. WordNet contains 7 types of relations between noun-synsets, but the main hierarchy in WordNet is built on hyponym relations, which are similar to subclass relations. W3C has translated WordNet version 2.0 into RDF/OWL[14].

The original WordNet model is a rich and well-designed model. However, some tools may have problems with the fact that the synsets are instances rather than classes. Therefore, for the purpose of this OAEI task, we have translated the hyponym hierarchy in a `skos:broader` hierarchy, making the synsets `skos:Concepts`.

DBpedia DBPedia contains 2.18 million resources or "things", each tied to an article in the English language Wikipedia. The "things" are described by titles and abstracts in English and often also in other languages, including Dutch. DBPedia "things" have numerous properties, such as categories, properties derived from the wikipedia 'infoboxes', links between pages within and outside wikipedia, etc. The purpose of this task is to map the DBPedia "things" to WordNet synsets and GTAA concepts.

[13] `http://wordnet.princeton.edu/`

[14] `http://www.w3.org/2006/03/wn/wn20/`

9.2 Evaluation Setup

We evaluate the results of the three alignments (GTAA-WordNet, GTAA-DBPedia, WordNet-DBPedia) in terms of precision and recall. We present measures for each GTAA facet separately, instead of a global value, because each facet could lead to very different performance.

In the precision and recall calculations, we use a kind of semantic distance; we take into account the distance between a correspondence that we find in the results and the ideal correspondence that we would expect for a certain concept. For each equivalence relation between two concepts in the results, we determine if (i) one is equivalent to the other, (ii) one is a broader/narrower concept than the other, (iii) one is in none of the above ways related to the other. In case (i) the correspondence counts as 1, in case (ii) the correspondence counts as 0.5 and in case (iii) as 0.

Precision We take samples of 100 correspondences per GTAA facet for both the GTAA-DBPedia and the GTAA-WordNet alignments and evaluate their correctness in terms of exact match, broader, narrower or related match, or no match. The alignment between WordNet and DBPedia is evaluated by inspection of a random sample of 100 correspondences.

Recall Due to time constraints, we only determine recall of two of the four GTAA facets: People and Subjects. These are the most extreme cases in terms of size and precision values. We create a small reference alignment from a random sample of 100 GTAA concepts per facet, which we manually map to WordNet and DBPedia. The result of the GTAA-WordNet and GTAA-DBPedia alignments are compared to the reference alignments. We do not provide a recall measure for the DBPedia-WordNet correspondence.

9.3 Results

Only one participant, DSSim, participated in the VLCR task. The evaluation of the results therefore focuses on the differences between the three alignments, and the four facets of the GTAA. Table 15 shows the number of concepts in each resource and the number of correspondences returned for each resource pair. The largest number of correspondences was found between DBpedia and WordNet (28,974), followed by GTAA-DBPedia (13,156) and finally GTAA-WordNet (2,405). We hypothesize that the low number of the latter pair is due to the multilingual nature. Except for 9 concepts, all GTAA concepts that were mapped to DBPedia were also mapped to WordNet.

Precision The precision of the GTAA-DBPedia alignment is higher than that of the GTAA-WordNet alignment. A possible explanation is the high number of disambiguation errors for WordNet, which is much finer grained than for GTAA or DBPedia.

A remarkable difference can be seen in the People facet. It is the worst scoring facet in the GTAA-WordNet alignment (10%), while it is the best facet in GTAA-DBPedia (94%). Inspection of the results revealed what caused the many mistakes for WordNet: almost none of the people in GTAA are present in WordNet. Instead of giving up, DSSim continues to look for a correspondence and maps the GTAA person to a lexically similar word in WordNet. This problem is apparently not present in DBPedia. Although we do not yet fully understand why not, an important factor is that more Dutch people are represented in DBPedia.

Vocabulary		#concepts	#corr to WN	#corr to DBP	#corr to GTAA
Wordnet		82.000	n.a.	28974	2405
DBPedia		2180.000	28974	n.a.	13156
GTAA		160.000	2405	13156	n.a.
Facet:	Subject	3800	655	1363	n.a.
	Person	97.000	82	2238	n.a.
	Name	27.000	681	3989	n.a.
	Location	14.000	987	5566	n.a.

Table 15. Number of correspondences in each alignment.

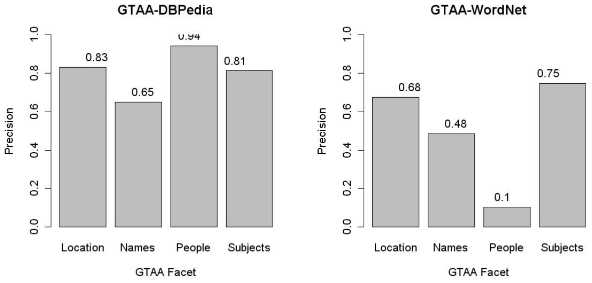

Fig. 11. Estimated precision of the alignment between GTAA and DBpedia (left) and WordNet (right).

Apart from the People facet, the differences between the facets are consistent over the GTAA-DBPedia and GTAA-WordNet alignments. Subjects and Locations score high, Names somewhat less.

The alignment between DBPedia and WordNet had a precision of 45%. DBPedia contains type links (wordnet-type and `rdf:type`) to WordNet synsets. There was no overlap between the alignment submitted by DSSim and these existing links.

Recall We created reference alignments by matching samples of 100 concepts from the People and Subjects facets to both DBPedia and WordNet. However, none of the People in our sample of 100 GTAA People could be mapped to WordNet. Therefore, recall for this particular alignment could not be detemined.

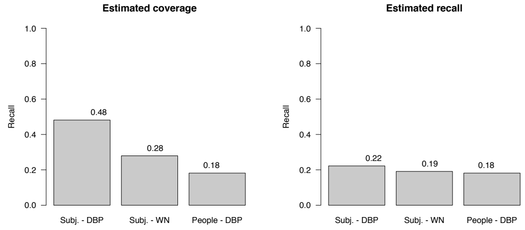

Fig. 12. Estimated coverage (left) and recall (right) for the alignments between the Subject facet of GTAA and DBpedia and WordNet, and for the alignment between the People facet of GTAA and DBpedia.

Figure 12 shows how many of the GTAA Subject and People in our reference alignment were also found by DSSim. We call this *coverage*. The second figure depicts how many GTAA concept in our reference alignment were found by DSSim to the exact same DBPedia/WordNet concept, which is the conventional definition of recall. All three alignments had a similar recall score of aroud 20%.

9.4 Summary of the results

Tables 16 and 17 summarize the result.

	Precision			
Alignment	Subjects	People	Location	Names
GTAA-DBPedia	0.81 (11.6%)	0.94 (7.02%)	0.83 (11.1%)	0.65 (14.1%)
GTAA-WordNet	0.75 (12.8%)	0.1 (8.8%)	0.68 (13.8%)	0.48 (14.7%)

Table 16. Summary of the participant's precision scores (numbers in parentheses represent the different error margins).

	Recall		Estimated coverage	
Alignment	Subjects	People	Subjects	People
GTAA-DBPedia	0.22 (12.2%)	0.18 (11.3%)	0.48 (14.7%)	0.18 (11.3%)
GTAA-WordNet	0.19 (11.6%)	NA	0.28 (13.2%)	NA

Table 17. Summary of the participant's estimated recall and coverage scores (numbers in parentheses represent the different error margins).

9.5 Discussion

Other types of correspondence relations The VLCR task once more confirmed what was already known: more correspondence types are necessary than only exact matches. While inspecting alignments, we found many cases where a link between two concepts seems useful for a number of applications, without being equivalent. For example:

```
Subject:pausbezoeken[15]
  and List_of_pastoral_visits_of_Pope_John_Paul_II_outside_Italy.
Location:Venezuela and synset-Venezuelan-noun-1
Subject:Verdedigingswerken[16] and fortification
```

Using context When looking at the types of mistakes that were made, it became clear that a number of them could have been avoided by using the specific structure of the resources being matched. The fact that the GTAA is organized in facets, for example, can be used to disambiguate terms that appear both as a person and as a location. This information is represented by the `skos:inScheme` property. Examples of incorrect correspondences that might have been avoided if facet information was used are:

```
Person:GoghVincentvan -> synset-vacationing-noun-1
Location:Harlem -> synset-hammer-noun-8
Location:Melbourne -> synset-Melbourne-noun-1[17]
```

Another example of resource-specific structure that could help matching are the redirects between pages in Wikipedia or between "things" in DBPedia. DBPedia contains things for which no other information is available than a 'redirect' property pointing to another thing. The wikipedia page for "Gordon Summer" for example, is immediately referred to the page for "Sting, the musician". The titles of these referring pages could well serve as alternative labels, and thus aid the correspondence between the gtaa concept person:SummerGordon and the dbepdia thing Sting(musician).

Of course, there is a trade-off between the amount of resource-specific features that are taken into account and the general applicability of the matcher. However, some of the features discussed above, such as facet information, are found in a wide range of thesauri and are therefore serious candidates for inclusion in a tool.

Reflection on the evaluation Deciding which synset or DBpedia thing is the most suitable match for a GTAA concept is a non-trivial task, even for a human evaluator.

[15] Pope visits, in English.

[16] Defenses, in English.

[17] This synset indeed refers to "a resort town in east central Florida".

Often, multiple correspondences are reasonable. Therefore, the recall figures that are based on a hand-made reference alignment give a possibly too negative impression of the quality of the alignment. The evaluation task was further complicated because of the 'related' matches. There is a lack of clear definitions of when two concepts are related.

Another factor that has to be considered when interpreting the precision and recall figures, is the number of Dutch-specific concepts in the GTAA. For example, the concept Name:Diogenes denotes a Dutch TV program instead of the ancient Greek. Although the fact that Diogenes is in the Name facet and not in the People facet provides a clue of its intended meaning, it could be argued that this type of Dutch-specific concepts pose an unfair challenge to matchers.

During the evaluation process, we found cases in which DSSim mapped to a DBPedia disambiguation page instead of an actual article. We consider this to be incorrect, since it leaves the disambiguation task to the user.

10 Conference

The conference track involves matching several ontologies from the conference organization domain. Participant results have been evaluated along different modalities and a consensus workshop aiming at studying the elaboration of consensus when establishing reference alignments has been organised.

10.1 Test set

The collection consists of fifteen ontologies in the domain of organizing conferences. Ontologies have been developed within the OntoFarm project[18]. In contrast to last year's conference track, there is one new ontology and several new methods of evaluation.

The main features of this data 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 concepts with enough erudition.
- *Independence of ontologies.* Ontologies were developed independently and based on different resources, they thus capture the issues in organizing conferences from different points of view and with different terminologies.
- *Relative richness in axioms.* Most ontologies were equipped with description logic axioms of various kinds, which opens a way to use semantic matchers.

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

Participants had to provide either complete alignments or interesting correspondences (nuggets), for all or some pairs of ontologies. Participants could also take part in two different tasks. First, participants could find correspondences without any specific

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

application context given (generic correspondences). Second, participants could find out correspondences with regard to an application scenario: *transformation application*. This means that final correspondences are to be used for conference data transformation from one software tool for organizing conference to another one.

This year, results of participants were evaluated by five different methods: evaluation based on manual labeling, reference alignments, data mining method, logical reasoning, and on consensus of experts.

10.2 Evaluation and results

We had three participants. All of them delivered generic correspondences. Aside from results from evaluation methods (sections below) we deliver some simple observations about participants:

- DSSim and Lily delivered in total 105 alignments. All ontologies were matched to each other. ASMOV delivered 75 alignments. For our evaluation we do not consider alignments in which ontologies were matched to themselves.
- Two participants delivered correspondences with certainty factors between 0 and 1 (ASMOV and Lily); one (DSSim) delivered correspondences with confidence measures 0 or 1, where 0 is used to describe a correspondence as negative.
- DSSim and Lily delivered only equivalence, e.g., no subsumption, relations, while ASMOV also provided subsumption relations[19].
- All participants delivered class-to-class correspondences and property-to-property correspondences.

Evaluation based on manual labeling This kind of evaluation is based on sampling and manual labeling of random samples of correspondences because the number of all distinct correspondences is quite high. Particularly, we followed the method of *Stratified random sampling* described in [20]. Correspondences of each participant were divided into three subpopulations (strata) according to confidence measures[20]. For each stratum we randomly chose 75 correspondences in order to have 225 correspondences for manual labeling for each system; except the one stratum of the DSSim system with 150 correspondences.

In Table 18 there are data for each stratum and system where *Nh* is the size of the stratum, *nh* is the number of sample correspondences from the stratum, *TP* is the number of correct correspondences from sample from the stratum, and *Ph* is an approximation of precision for the correspondences in the stratum. Furthermore, based on the assumption that this adheres to *binomial distribution* we computed *margin of errors* (with confidence of 95%) for the approximated precision for each system based on equations from [20]. In Table 19 there are measures for the entire populations. We computed approximated precision $P*$ in the entire population as weighted average from the approximated precisions of each strata. Finally, we also computed so-called 'relative'

[19] Finally, no current evaluation methods did take into account subsumption correspondences. Considering these correspondences in evaluation methods is our plan for next year of the conference track.

[20] DSSim provided merely 'certain' correspondences, so there is just one stratum for this system.

	(0,0.3]		(0.3,0.6]		(0.6,1.0]		
system	ASMOV	Lily	ASMOV	Lily	ASMOV	Lily	DSSim
Nh	779	426	349	911	135	407	1950
nh	75	75	75	75	75	75	150
TP	16	33	38	27	51	39	46
Ph	21%	44%	51%	36%	68%	52%	30%
	±12%	±12%	±12%	±12%	±12%	±12%	±8%

Table 18. Summary of the results for samples.

	ASMOV	DSSim	Lily
P*	34% ± 10%	30% ± 8%	42% ± 10%
rrecall	18%	14%	17%

Table 19. Summary of the results for entire populations.

recall (*rrecall*) that is computed as ratio of the number of all correct correspondences (sum of all correct correspondences per one system) to the number of all correct correspondences found by any of systems (per all systems). This relative recall was computed over stratified random samples, so it is rather sample relative recall.

Discussion Although the ASMOV system achieves the highest result in two strata and the Lily system in the approximated precision P*, because of overlapping margins of errors we cannot say that a system outperforms another. In order to make approximated results more decisive we should take larger samples. Regarding relative recall, ASMOV achieves the highest value.

Evaluation based on reference alignments This is the classical evaluation method where the alignments from participants are compared against the *reference alignment*. So far we have built the reference alignment over five ontologies (cmt, confOf, ekaw, iasted, sigkdd, i.e. 10 alignments); we plan to cover the whole collection in the future. The decision about each correspondence was based on majority vote of three evaluators. In the case of disagreement among evaluators, the given correspondence was the subject of broader public discussion during the Consensus building workshop in order to find consensus and update the reference alignment, see the section (below) about the Evaluation based on the consensus of experts.

	t=0.2			t=0.5			t=0.7		
	P	R	F-meas	P	R	F-meas	P	R	F-meas
ASMOV	51.8%	38.6%	44.2%	72.2%	11.4%	19.7%	100.0%	6.1%	11.6%
DSSim	34.0%	57.9%	42.9%	34.0%	57.9%	42.9%	34.0%	57.9%	42.9%
Lily	43.2%	50.0%	46.3%	60.4%	28.1%	38.3%	66.7%	8.8%	15.5%

Table 20. Recall, precision and F-measure for three different thresholds

In Table 20, there are traditional *precision* (P), *recall* (R), and *F-measure* (F-meas) computed for three diverse thresholds (0.2, 0.5, and 0.7). As we have mentioned, these results are biased because the current reference alignment only covers a subset of all ontology pairs from the OntoFarm collection.

Discussion All systems achieve the highest F-measure for threshold 0.2, while the Lily system has the highest F-measure of 46.3%. The ASMOV system achieves the highest precision for each of three thresholds (51.8%, 72.2%, 100%) however it is at the expense of recall that is the lowest for each of three thresholds (38.6%, 11.4%, 6.1%). The highest recall (57.9%) was obtained by the DSSim system.

Evaluation based on data mining method This kind of evaluation is based on data mining, and the goal is to reveal non-trivial findings about the participating systems. These findings relate to the relationships between the particular system and features such as the confidence measure, validity, kinds of ontologies, particular ontologies, and *mapping patterns*. Mapping patterns have been introduced in [19]. For the purpose of our current experiment we extended detected mapping patterns with some patterns inspired by *correspondence patterns* [16] and with *error mapping patterns*.

Basically, mapping patterns are patterns dealing with (at least) two ontologies. These patterns reflect the *the structure of ontologies* on the one side, and on the other side they include correspondences between entities of ontologies. Initially, we discover some mapping patterns such as occurrences of some complex structures in the participants results. They are neither the result of a deliberate activity of humans, nor they are a priori 'desirable' or 'undesirable'. Here are three such mapping patterns between concepts:

- MP1 (Parent-child triangle): it consists of an equivalence correspondence between A and B and an equivalence correspondence between A and a child of B, where A and B are from different ontologies.
- MP2 (Mapping along taxonomy): it consists of simultaneous equivalence correspondences between parents and between children.
- MP3 (Sibling-sibling triangle): it consists of simultaneous correspondences between class A and two sibling classes C and D where A is from one ontology and C and D are from another ontology.

This year, we added three mapping patterns inspired by correspondence patterns [16]:

- MP4: it is inspired by the 'class by attribute' correspondence pattern, where the class in one ontology is restricted to only those instances having a particular value for a a given attribute/relation.
- MP5: it is inspired by the 'composite' correspondence pattern. It consists of a class-to-class equivalence correspondence and a property-to-property equivalence correspondence, where classes from the first correspondence are in the domain or in the range of properties from the second correspondence.
- MP6: it is inspired by the 'attribute to relation' correspondence pattern where a datatype and an object property are aligned as an equivalence correspondence.

Furthermore, there are error mapping patterns, which can disclose incorrect correspondences:

- MP7: it is the variant of MP5 ‘composite pattern’. It consists of an equivalence correspondence between two classes and an equivalence correspondence between two properties, where one class from the first correspondence is in the domain and another class from that correspondence is in the range of equivalent properties, except the case where domain and range is the same class.
- MP8: it consists of an equivalence correspondence between A and B and an equivalence correspondence between a child of A and a parent of B where A and B are from different ontologies. It is sometimes reffered to as criss-cross pattern.
- MP9: it is the variant of MP3, where the two sibling classes C and D are disjoint.

	MP1	MP2	MP3	MP4	MP5	MP6	MP7	MP8	MP9
ALL	0/543/0	255/146/115	0/527/0	261/828/354	467/115/585	132/115/151	0/6/13	0/7/4	0/165/0
REF	0/70/0	39/19/17	0/58/0	35/88/35	51/6/29	1/2/3	0/0/0	0/3/0	0/27/0

Table 21. Occurrences of mapping patterns in participants results.

In Table 21 there are numbers of correspondences found by each system (ASMOV/DSSim/Lily) that belong to a particular mapping pattern. The row ‘ALL’ relates to all equivalence correspondences delivered by participants with confidence measure higher than 0.0 (1540/1950/1744). The row ‘REF’ relates to all equivalence correspondences delivered by participants with confidence measure higher than 0.0 for pairs of ontologies for which there exists the reference alignment (182/194/132).

For the *data-mining analysis* we employed the *4ft-Miner* procedure of the *LISp-Miner* data mining system[21] for mining of *association rules*. For the sake of brevity we mention a few examples of interesting *association hypotheses* discovered[22]:

- In correspondences with low confidence measure [0,0.4) the ASMOV system comes 1.2 times more often with incorrect correspondences for *cmt* and *confOf* pair of ontologies than all systems with such (incorrect) correspondences for those two ontologies with all confidence measures (on average).
- The Lily system outputs almost three times more often correspondences that belong to the mapping pattern MP7 than do all systems (on average).
- In correspondences with low confidence measure [0,0.4) the Lily system comes 1.2 times more often with correct correspondences for pairs of ontologies with *iasted* ontology than all systems with such (correct) correspondences for those pairs of ontologies with all confidence measures (on average).

Discussion The abovementioned hypotheses disclose potentially interesting relationships for the developers of systems. By Table 21 (particularly numbers for MP7, MP8, and mainly for MP9) we could say that application of error mapping patterns

[21] http://lispminer.vse.cz/

[22] For association hypotheses with confidence measures we used REF correspondences, otherwise we used ALL correspondences.

would improve the systems' performance (for Lily to some degree and especially for DSSim) in terms of precision, while the results of the ASMOV system do not contain any instances of error mapping patterns due to its semantic verification phase.

Evaluation based on alignment incoherence Several ways to measure the incoherence of an alignment have been proposed in [13]. In the following we focus on the maximum cardinality measure m^t_{card} which has been introduced as revision based measure. The m^t_{card} measure compares the number of correspondences which have to be removed to arrive at a coherent subset with the number of all correspondences in the alignment. The conference ontologies are well suited for an analysis of alignment incoherence since most of them contain negation as well as different kinds of restrictions exploiting the range of OWL-DL expressivity.

Due to practical considerations we decided to modify the approach with respect to two aspects. First, we observed that many logical problems induced by an alignment are related to properties. Therefore, we applied a different definition of incoherence taking property unsatisfiability into account. We defined an ontology to be incoherent whenever there exists an unsatisfiable concept or property. This extends the classical approach in which ontology incoherence depends only on the unsatisfiability of concepts (see for example [14]). Second, we observed that matching object properties on datatype properties might be an appropriate way to cope with semantic heterogeneity. Nevertheless, such a correspondence would directly result in an incoherent alignment based on the direct natural translation of a correspondence as axiom. Therefore, we used a slightly modified variant of the natural translation and translated each correspondence between properties R_1 and R_2 into an axiom $\exists R_1.\top \equiv \exists R_2.\top$ (we only considered equivalence correspondences).

System	Alignments	Coherent	Mean	Median
ASMOV	44 (1010)	8	0.135	0.14
Lily	45 (851)	9	0.138	0.145
DSSim	45 (769)	3	0.206	0.166

Table 22. Number of evaluated alignments (and total of correspondences), number of coherent alignments, mean and median for the maximum cardinality measure..

In our experimental evaluation we considered only a subset of 10 ontologies and evaluated the alignments between all possible pairs. We excluded five ontologies (Cocus, Confious, Iasted, Paperdyne and OpenConf) because we only focused on alignments submitted by each participant and encountered reasoning problems for some of these ontologies. Table 22 summarizes the main results. First of all we notice that only a small fraction of submitted alignments is coherent. For ASMOV and Lily 18% resp. 20% of the evaluated alignments were coherent, while DSSim generated only 7% coherent alignments. We also computed the mean of the m^t_{card} measure over all analyzed alignments. We observe that ASMOV and Lily generate alignments with a lower degree of incoherence (0.135 and 0.138) compared to DSSim (0.206).

The distribution of measured values additionally supports our first impression. Figure13 shows the second and third quartile as well as the median of the values measured via m^t_{card}. While Lily and especially ASMOV found a way to prevent highly incoherent alignments, 25% of the alignments generated by DSSim have a degree of incoherence greater or equal than 0.288. For each of these alignments there are logical reasons to remove at least one-fourth of its correspondences. The differences between ASMOV, Lily and DSSim revealed by our incoherence analysis fits with the differences we reported on the occurence of the error mapping patterns MP7 to MP9.

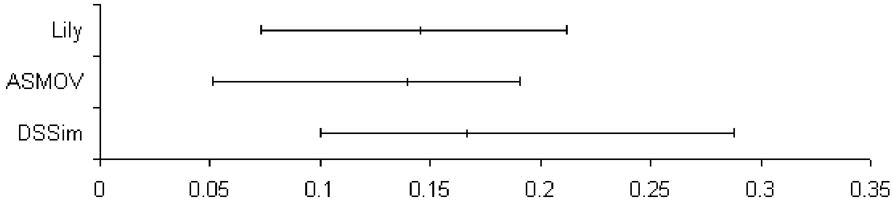

Fig. 13. Distribution of m^t_{card} values, depicting second quartile, median, and third quartile.

Discussion Some of the participants implemented a component to debug or validate generated alignments, namely ASMOV and Lily. To our knowledge these debugging techniques are based on detecting certain structural patterns in correspondence pairs (MP7 to MP9 can be seen as examples of such patterns). Although these strategies cannot ensure the coherence of an alignment, such an approach is nevertheless an efficient way to avoid full-fledged reasoning while increasing the degree of coherence. Taking alignment coherence into account can be a useful guide for improving the results of a matching system and our results suggest that there is still room for improvement.

Evaluation based on consensus of experts During so-called Consensus building workshop we discussed 5 controversial correspondences. The main goal of this discussion among experts was to find consensus about those correspondences and track arguments against and favour. This session ratified insights from previous years and disclosed that finding consensus is time-consuming and not an easy activity however doable. Some other relevant topics were raised. For instance, open-world assumption vs. closed-world assumption was considered as an important factor for understanding the description of entities in ontologies. The need for expressive alignments also arouse for expressing complex correspondences combining several elements (classes or properties). The reached consensus is captured in the reference alignment and discussion can be further proceed in the blog[23].

[23] http://keg.vse.cz/oaei/

10.3 Conclusion

In conclusion, we evaluated participant results from diverse perspectives via five distinct evaluation methods. For next year of this track, we also plan to evaluate subsumption correspondences and further extend the reference alignment. Based on the participants' feedback we changed ontologies from the OntoFarm collection in order to be OWL DL compliant for the next year of the conference track.

11 Lesson learned and suggestions

The lessons learned for this year are relatively similar to those of previous years. But there remain lessons not really taken into account that we identify with an asterisk (*). We reiterate those lessons that still apply with new ones:

A) Unfortunately, we have not been able to maintain the better schedule of last year. With the schedule reduced by one month (thus in overall having about 3 months), it is very difficult to run OAEI.

B) Some of the best systems of last year did not enter. The invoked reasons were: not enough time and/or no improvement in the systems. This pleads for continous instead of yearly evaluation.

C) The trend that there are more matching systems able to enter such an evaluation seems to slow down. However, the number of tracks the existing systems are able to consider still very encouraging for the progress of the field.

D) We can confirm that systems that enter the campaign for several times tend to improve over years.

E*) The benchmark test case is not discriminant enough between systems. It is still useful for evaluating the strengths and weaknesses of algorithms but does not seem to be sufficient anymore for comparing algorithms. We have improved tests this year, while preserving comparability with previous years, but more is required, in particular in automatic test generation.

F) We have had more proposals for test cases this year. 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 becomes easier to perform the evaluation.

G) There are now test cases where non equivalence-only alignments matter and there are systems, e.g., ASMOV, Aroma, TaxoMap, which are able to deliver such alignments. We thus intent to have such a test case next year. The discussion about instance matching tests has also aroused.

H) The robustness of evaluation tools make that, like last year, we had very few syntactic problems this year. However, it seems that many matchers are too dependent on particular operating systems and still many ones do not deal correctly with ontology URIs (see the Error cells in Table 3).

I) The partition between systems able to deal with large ontologies and systems unable to do it seems to be transforming gradually: systems seem to be able to perform more tasks. However, this requires an important amount of manpower.

12 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 with 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, see [17] for a detailed discussion on the ontology matching challenges.

13 Conclusions

This year we had less systems overall entering the evaluation campaign with still a significant number of systems. It seems however that they entered more tests individually (50 last year overall against 48 this year), so systems seem to be more up to the challenge.

As noticed the previous years, systems which do not enter for the first time are those which perform better. This shows that, as expected, the field of ontology matching is getting stronger (and we hope that evaluation has been contributing to this progress).

All participants have provided description of their systems and their experience in the evaluation. These OAEI papers, like the present one, have not been peer reviewed. However, they are full contributions to this evaluation exercise and reflect the hard work and clever insight people put in the development of participating systems. 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. Sometimes participants offer alternate evaluation results.

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

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

Acknowledgments

We warmly thank each participant of this 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 Martin Ringwald and Terry Hayamizu for providing the reference alignment for the anatomy ontologies.

Thanks to Andrew Bagdanov, Aureliano Gentile, Gudrun Johannsen (Food and Agriculture Organization of the United Nations) for evaluating the FAO task. We also thank the teams of Agricultural Organization of the United Nations (FAO) for allowing us to use their ontologies. Caterina Caraciolo and Jérôme Euzenat have been partially supported by the European integrated project NeOn (IST-2005-027595).

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.

We thank Chris Bizer, Fabian Suchanec and Jens Lehman for their help with the DBPedia dataset. We also thank Willem van Hage for his advices. We gratefully acknowledge the Dutch Institute for Sound and Vision for allowing us to use the GTAA.

We are grateful to Peter Bartoš (Brno University of Technology, CZ) for participating in creation of partial reference alignment for the conference track. In addition, Ondřej Šváb-Zamazal and Vojtěch Svátek were supported by the IGA VSE grant no.20/08 “Evaluation and matching ontologies via patterns”.

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).

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 international 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 workshop on Integrating Ontologies*, Banff (CA), 2005.
3. 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.
4. Marc Ehrig and Jérôme Euzenat. Relaxed precision and recall for ontology matching. In *Proceedings of the K-Cap workshop on Integrating Ontologies*, pages 25–32, Banff (CA), 2005.
5. Jérôme Euzenat. An API for ontology alignment. In *Proceedings of the 3rd International Semantic Web Conference (ISWC)*, pages 698–712, Hiroshima (JP), 2004.
6. 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 international workshop on Ontology Matching, Athens (GA US)*, pages 73–95, 2006.
7. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer, Heidelberg (DE), 2007.

8. Jérôme Euzenat, Antoine Isaac, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Svab, Vojtech Svatek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative 2007. In Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, and Bin He, editors, *Proceedings of the 2nd ISWC international workshop on Ontology Matching, Busan (KR)*, pages 96–132, 2007.
9. Fausto Giunchiglia, Mikalai Yatskevich, Paolo Avesani, and Pavel Shvaiko. A large scale dataset for the evaluation of ontology matching systems. *The Knowledge Engineering Review Journal*, (24(2)), 2009, to appear.
10. Ryutaro Ichise, Masahiro Hamasaki, and Hideaki Takeda. Discovering relationships among catalogs. In *Proceedings of the 7th International Conference on Discovery Science*, pages 371–379, Padova (IT), 2004.
11. Ryutaro Ichise, Hideaki Takeda, and Shinichi Honiden. Integrating multiple internet directories by instance-based learning. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 22–28, Acapulco (MX), 2003.
12. Antoine Isaac, Henk Matthezing, Lourens van der Meij, Stefan Schlobach, Shenghui Wang, and Claus Zinn. Putting ontology alignment in context: Usage scenarios, deployment and evaluation in a library case. In *Proceedings of the 5th European Semantic Web Conference (ESWC)*, pages 402–417, Tenerife (ES), 2008.
13. Christian Meilicke and Heiner Stuckenschmidt. Incoherence as a basis for measuring the quality of ontology mappings. In *Proceedings of the 3rd ISWC international workshop on Ontology Matching*, pages 1–12, Karlsruhe (DE), 2008.
14. Guilin Qi and Anthony Hunter. Measuring incoherence in description logic-based ontologies. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, pages 381–394, Busan (KR), 2007.
15. Marta Sabou, Mathieu d'Aquin, and Enrico Motta. Using the semantic web as background knowledge for ontology mapping. In *Proceedings of the ISWC international workshop on Ontology Matching*, pages 1–12, Athens (GA US), 2006.
16. Francois Scharffe and Dieter Fensel. Correspondence patterns for ontology alignment. In *Proceedings of the 16th International Conference on Knowledge Acquisition, Modeling and Management (EKAW)*, pages 83–92, Acitrezza (IT), 2008.
17. Pavel Shvaiko and Jérôme Euzenat. Ten challenges for ontology matching. In *Proceedings of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)*, pages 1164–1182, Monterrey (MX), 2008.
18. 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.
19. Ondrej Svab, Vojtech Svatek, and Heiner Stuckenschmidt. A study in empirical and 'casuistic' analysis of ontology mapping results. In *Proceedings of the 4th European Semantic Web Conference (ESWC)*, pages 655–669, Innsbruck (AU), 2007.
20. Willem Robert van Hage, Antoine Isaac, and Aleksovski, Zharko. Sample evaluation of ontology matching systems. In *Proceedings of the ISWC workshop on Evaluation of Ontologies and Ontology-based tools*, pages 41–50, Busan (KR), 2007.

Roma, Grenoble, Tokyo, Amsterdam, Trento, Mannheim, and Prague, December 2008