First Results of the Ontology Alignment Evaluation Initiative 2010*

Jérôme Euzenat[1], Alfio Ferrara[6], Christian Meilicke[2], Juan Pane[3], François Scharffe[1], Pavel Shvaiko[4], Heiner Stuckenschmidt[2], Ondřej Šváb-Zamazal[5], Vojtěch Svátek[5], and Cássia Trojahn[1]

[1] INRIA & LIG, Montbonnot, France
{jerome.euzenat,francois.scharffe,cassia.trojahn}@inrialpes.fr
[2] University of Mannheim, Mannheim, Germany
{christian,heiner}@informatik.uni-mannheim.de
[3] University of Trento, Povo, Trento, Italy
pane@dit.unitn.it
[4] TasLab, Informatica Trentina, Trento, Italy
pavel.shvaiko@infotn.it
[5] University of Economics, Prague, Czech Republic
{svabo,svatek}@vse.cz
[6] Universita degli studi di Milano, Italy
ferrara@dico.unimi.it

Abstract. Ontology matching consists of finding correspondences between entities of two ontologies. OAEI campaigns aim at comparing ontology matching systems on precisely defined test cases. Test cases can use ontologies of different nature (from simple directories to expressive OWL ontologies) and use different modalities, e.g., blind evaluation, open evaluation, consensus. OAEI-2010 builds over previous campaigns by having 4 tracks with 6 test cases followed by 15 participants. This year, the OAEI campaign introduces a new evaluation modality in association with the SEALS project. A subset of OAEI test cases is included in this new modality. The aim is to provide more automation to the evaluation and more direct feedback to the participants. This paper is an overall presentation of the OAEI 2010 campaign.

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 [9]. The main goal of OAEI is to compare systems and algorithms on the same basis and to allow anyone for drawing conclusions about the best matching strategies.

* This is only a preliminary and incomplete version of the paper. It presents a partial and early view of the results. The final results will be published on the OAEI web site shortly after the ISWC 2010 workshop on Ontology Matching (OM-2010) and will be the only official results of the campaign.

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

Our ambition is that from such evaluations, tool developers can learn and improve their systems.

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) [13]. Then, unique OAEI campaigns occurred in 2005 at the workshop on Integrating Ontologies held in conjunction with the International Conference on Knowledge Capture (K-Cap) [1], in 2006 at the first Ontology Matching workshop collocated with ISWC [8], in 2007 at the second Ontology Matching workshop collocated with ISWC+ASWC [7], in 2008, OAEI results were presented at the third Ontology Matching workshop collocated with ISWC [3], and in 2009, OAEI results were presented at the fourth Ontology Matching workshop collocated with ISWC [6], Finally, in 2010, OAEI results are presented at the fifth Ontology Matching workshop collocated with ISWC, in Shanghai, China[2].

The main innovation of this year is the adoption of an environment for automatically processing evaluations (§2.2), which has been developed in coordination with the SEALS project[3]. This project aims at providing standardized datasets, evaluation campaigns for typical semantic web tools, including ontology matching, and a software infrastructure for automatically executing evaluations. This year, a subset of OAEI datasets is included in the SEALS modality. The goal is to provide better direct feedback to the participants and a more common ground to the evaluation.

We have discontinued the oriented alignment track of last year because there was not enough organisational resources to guarantee a satisfying evaluation. The instance track has been maintained.

This paper serves as an introduction to the evaluation campaign of 2010 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 evaluation methodology that has been used. Sections 3-7 discuss in turn the settings and the results of each of the test cases. Section 8 overviews lessons learned from the campaign. Finally, Section 9 outlines future plans and Section 10 concludes the paper.

2 General methodology

We first present the test cases proposed this year to OAEI participants. Then, we present the evaluation environment, which has been used for participants to test their systems and launch their evaluation experiments for the campaign. Next, we describe the steps of the OAEI campaign and report on the general execution of the campaign. In particular, we list participants and the tests they have considered.

[2] http://om2010.ontologymatching.org
[3] http://www.seals-project.eu

2.1 Tracks and test cases

This year's campaign has consisted of 4 tracks gathering 6 data sets and different evaluation modalities:

The benchmark track (§3): Like in previous campaigns, a systematic benchmark series has been proposed. The goal of this benchmark series is to identify the areas in which each alignment 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 capabilities:

- **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.
- **Conference (§5):** The goal of this track is to find all correct correspondences within a collection of ontologies describing the domain of organising conferences (the domain being well understandable for every researcher). Additionally, 'interesting correspondences' are also welcome. Results will be evaluated automatically against a reference alignment and by data-mining and logical reasoning techniques. Sample of correspondences and 'interesting correspondences' will be evaluated manually.

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

- **Directory (§6):** The directory real world case consists of matching web site directories (like open directory or Yahoo's). This year the track consists of two modalities, the first is composed by more than 4 thousand elementary tests, and the second is composed by a single test which matches two big directories (2854 and 6555 nodes each).

Instance matching (§7): The instance data matching track aims at evaluating tools able to identify similar instances among different RDF and OWL datasets. It features Web datasets, as well as a generated benchmark.

- **IMEI** This task (imei) is focused on RDF and OWL data in the context of the Semantic Web. Participants were asked to execute their algorithms against various datasets and their results were evaluated by comparing them with a pre-defined reference alignment. Results are evaluated according to standard precision and recall metrics.
- **Very large crosslingual resources:** The purpose of this task (vlcr) is (1) to create alignments between large thesauri in different languages and (2) to align these thesauri to other sources on the Linked Data Web. This is seen as a step towards opening up and connecting large collections of data all around the world. In the vlcr task we align three resources to each other: the Thesaurus of the Netherlands Institute for Sound and Vision (called GTAA), the New York Times subject headings and DBpedia.

The datasets Benchmark, Anatomy and Conference have been evaluated using the SEALS service. The reason for this is twofold: on the one hand these data sets are well

known to the organizers and have been used in many evaluations, contrary to the test cases of the instance data sets, for instance. On the other hand, these data sets come with a high quality reference alignment which allows for computing the compliance based measures, such as precision and recall.

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

This year again, we had to cancel a data set. The vlcr (Very large crosslingual resources) data set had not enough participants to be maintained.

test	formalism	relations	confidence	modalities	language
benchmarks	OWL	=	[0 1]	open	EN
anatomy	OWL	=	[0 1]	open	EN
conference	OWL-DL	=, <=	[0 1]	blind+open	EN
directory	OWL	=	1	blind+open	EN
ars	RDF	=	[0 1]	open	EN
tap	RDF	=	[0 1]	open	EN
iimb	RDF	=	[0 1]	open	EN
vlcr	SKOS +OWL	exact-, closeMatch	[0 1]	blind expert	DU+EN

Table 1. Characteristics of test cases (open evaluation is made with already published reference alignments and blind evaluation is made by organizers from reference alignments unknown to the participants).

2.2 The SEALS evaluation service

This year, participants have used the SEALS evaluation service for testing their systems and launching their own evaluation experiments. A first version of this evaluation service[4] is based on the use of a web service interface wrapping the functionality of a matching tool to be evaluated. Participants were invited to extend a web service interface[5] and deploy their matchers as web services, which are accessed during the evaluation process. This setting allows for participants debuging their systems, running their own evaluations and manipulating the results immediately in a direct feedback cycle.

In order to start an evaluation, the participant must specify the URL of the matcher service and the name of the matching system to be evaluated as well as he must select the data set to be used (Anatomy, Benchmark or Conference). Then, the specified web service is validated by the system (two simple ontologies are used to check if the matcher generate alignments in the correct format). In case of a problem, the concrete validation error is displayed to the user as direct feedback. In case of a successfully completed validation, the system returns a confirmation message and continues with the evaluation process. The values of precision, recall and F–measure are then displayed for each test case. The complete description of the preliminary version of the SEALS evaluation service for matching tools can be found in [14].

[4] `http://seals.inrialpes.fr/platform/`

[5] `http://alignapi.gforge.inria.fr/tutorial/tutorial5/`

Furthermore, organizers have a tool for accessing the results registered for the campaign as well as all evaluations being carried out in the evaluation service (even the evaluation executed for testing purposes). Manipulation of results can be done via an OLAP application (Figure 1).

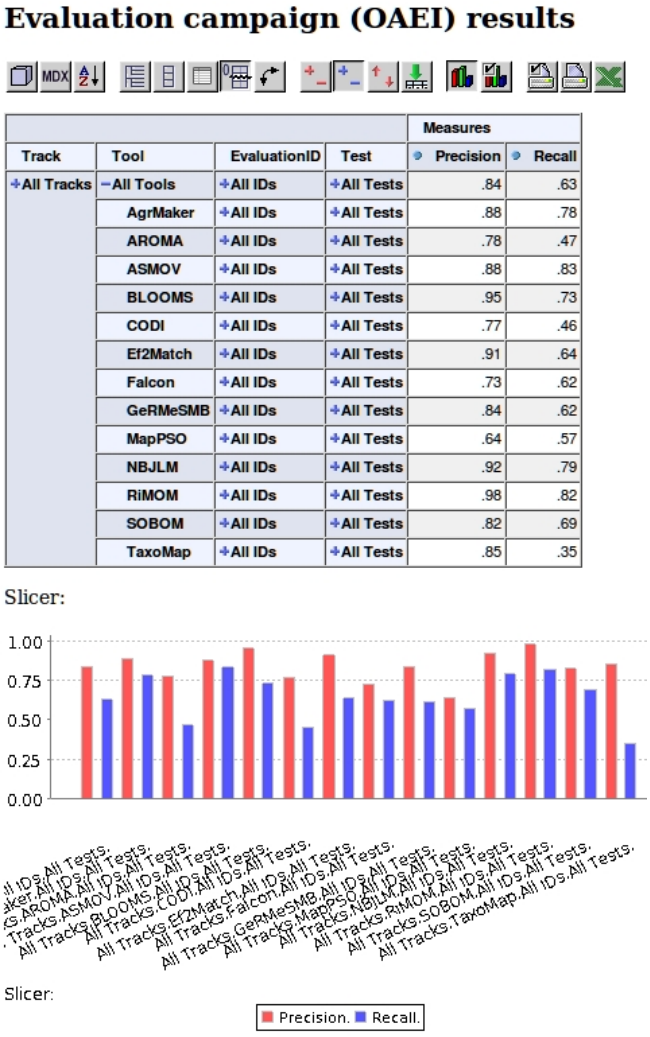

Fig. 1. Using OLAP for results visualization.

2.3 Preparatory phase

Ontologies to be matched and (where applicable) reference alignments have been provided in advance during the period between June 1^{st} and June 21^{st}, 2010. 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 8^{th}. The data sets did not evolve after this period.

2.4 Preliminary tests

In this phase, participants were invited to test their systems in order to ensure that the systems can load the ontologies to be aligned and generate the alignment in the correct format, the Alignment format expressed in RDF/XML [5]. Participants have been requested to provide (preliminary) results by August 30^{th}.

For the SEALS modality, testing could be conducted using the evaluation service while for the other tracks participants submitted their preliminary results to the organizers, who analyzed them semi-automatically, often detecting problems related to the format or to the naming of the required results files.

2.5 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, participants should not use the data (ontologies and reference alignments) from other test cases to help their algorithms. In most cases, ontologies are described in OWL-DL and serialized in the RDF/XML format.

For the standard OAEI modalities, participants had to run their systems on their own machines and submit the results via mail to the organizers. SEALS participants ran their systems via the SEALS evaluation service. They got a direct feedback on the results and could validate them as final results. Furthermore, SEALS participants were invited to register their tools by that time in the SEALS portal[6].

Participants also provided the papers that are published hereafter and a link to their systems and their configuration parameters.

2.6 Evaluation phase

In the evaluation phase, the organizers have evaluated the alignments provided by the participants and returned comparisons on these results. Final results were due by October 4^{th}, 2010. In the case of blind tests, only the organizers did the evaluation with regard to the withheld reference alignments.

Concerning SEALS, the participants have used the evaluation service for registering their results for the campaign. The evaluation effort is minimized due the fact that the results are automatically computed by the services in the evaluation service as well as organizers have an OLAP application for manipulating and visualizing the results.

[6] `http://www.seals-project.eu/join-the-community/`

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.

2.7 Comments on the execution

Since a few years, the number of participating systems has remained roughly stable: 4 participants in 2004, 7 in 2005, 10 in 2006, 17 in 2007, 13 in 2008, 16 in 2009 and 15 in 2010.

The number of covered runs has decreased more than expected: 37 in 2010, 53 in 2009, 50 in 2008, and 48 in 2007. This may be due to the increasing specialization of tests: some systems are specifically designed for instance matching or for anatomy.

This year many of the systems are validated through web services thanks to the SEALS evaluation service. For the next OAEI campaign, we expect to be able to actually run the matchers in a controlled evaluation environment, in order to test their portability and deployability. This will allows us for comparing system on a same execution basis.

The list of participants is summarized 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.

System	AgrMaker	AROMA	ASMOV	BLOOMS	CODI	Ef2Match	Falcon-AO	GeRMeSMB	LNR2	MapPSO	NBJLM	ObjectRef	RiMOM	SOBOM	TaxoMap	Total=15
Confidence	√	√	√	√		√	√	√	√	√	√	√	√	√	√	
benchmarks	√	√	√		√	√	√	√		√			√	√	√	11
anatomy	√		√	√	√	√		√			√			√	√	9
conference	√	√	√		√	√	√	√						√		8
directory			√					√		√					√	4
iimb			√		√				√			√	√			5
Total	3	2	5	1	4	3	2	4	1	2	1	1	2	3	3	37

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.

The set of participants is divided in two main categories: those who participated in the instance matching track and those who participated in ontology matching tracks.

Three systems (ASMOV, CODI, RiMOM) participated in both types of tracks. Last year only two systems (DSSim and RiMOM) has participated in both types of tracks.

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

3 Benchmark

The goal of the benchmark data set 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 data

The domain of this first test is Bibliographic references. It is based on a subjective view of what must be a bibliographic ontology. There may 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 one goal of these tests is to offer a permanent benchmark to be used by many, the test is an extension of the 2004 EON Ontology Alignment Contest, whose test numbering it (almost) fully preserves.

The tests are roughly the same as last year. 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

Eleven systems have participated in the benchmark track of this year's campaign (see Table 2). Four systems that had participated last year (AFlood, DSSim, Kosimap and Lily) did not participate this year, while two new systems (CODI and Ef2Match) have registered their results.

Table 3 shows the results, by groups of tests. For comparative purposes, the results of systems that have participated last year are also provided. 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.

As shown in Table 3, two systems are ahead: ASMOV and RiMOM, with AgrMaker as close follower, while SOBOM, GeRMeSMB and Ef2Match, respectively, had presented intermediary values of precision and recall. In the 2009 campaign, Lily and ASMOV were ahread, with aflood and RiMOM as followers, while GeRoME, AROMA, DSSim and AgrMaker had intermediary performance. The same group of best matchers has been presented in both campaigns. No system had strictly lower performance than edna.

Looking for each group of tests, in simple tests (1xx) all systems have similar performance, excluding TaxoMap which has presented low value of recall. As noted in previous campaigns, the algorithms have their best score with the 1xx test series. It is due the fact that there are no modifications in the labels of classes and properties in these tests and basically all matchers are able to deal with label similarity. For systematic tests (2xx), which allows better to distinguish the strengths of algorithms, ASMOV and RiMOM, respectively, are again ahead of the other systems, followed by AgrMaker, SOBOM, GeRMeSMB and Ef2Match, respectively, which have presented good performance, specially in terms of precision. Finally, for real cases (3xx), ASMOV (in average) provided the best results, with RiMOM and Ef2Match as followers. The best precision for these cases was obtained by the new participant CODI.

In general, the systems have improved their performance since last year: ASMOV and RiMOM improved their overall performance, AgrMaker and SOBOM have significantly improved their recall while MapPSO and GeRMeSBM improved precision. AROMA has significantly decreased in recall, for the three groups of tests. There is no unique set of systems ahead for all cases, what indicates that systems exploiting different features of ontologies perform accordingly to the features of each test cases.

As last year, the apparently best algorithms provide their results with confidence measures. It is thus possible to draw precision/recall graphs in order to compare them. Figure 2 shows the precision and recall graphs of this year. These results are only relevant for the results of participants who provide confidence measures different from 1 or 0 (see Table 2). Contrary to previous years these graphs are not drawn with the same principles as TREC's. They now show the real precision at n% recall and they stop when no more correspondences are available (then the end point corresponds to the

system	refalign		edna		AgrMaker		AROMA		ASMOV		CODI		Ef2Match		Falcon		GeRMeSMB		MapPSO		RiMOM		SOBOM		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.
2010																										
1xx	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.98	1.00	1.00	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	0.34
2xx	1.00	1.00	0.43	0.57	0.95	0.84	0.94	0.46	0.99	0.89	0.83	0.42	0.98	0.63	0.81	0.63	0.96	0.66	0.67	0.59	0.99	0.83	0.97	0.74	0.86	0.29
3xx	1.00	1.00	0.51	0.65	0.88	0.58	0.83	0.58	0.88	0.84	0.95	0.45	0.92	0.75	0.89	0.76	0.90	0.42	0.72	0.39	0.94	0.76	0.79	0.75	0.71	0.32
H-mean	1.00	1.00	0.45	0.58	0.95	0.84	0.94	0.48	0.99	0.89	0.84	0.44	0.98	0.65	0.82	0.65	0.96	0.67	0.68	0.60	0.99	0.84	0.97	0.75	0.86	0.29
2009																										
1xx	1.00	1.00	0.96	1.00	0.98	0.98	1.00	1.00	1.00	1.00							1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.97	1.00	0.34
2xx	1.00	1.00	0.41	0.56	0.98	0.60	0.98	0.69	0.96	0.85							0.92	0.71	0.73	0.73	0.93	0.81	0.97	0.46	0.90	0.23
3xx	1.00	1.00	0.47	0.82	0.92	0.79	0.85	0.78	0.81	0.82							0.68	0.60	0.54	0.29	0.81	0.82	0.92	0.55	0.77	0.31
H-mean	1.00	1.00	0.43	0.59	0.99	0.62	0.94	0.69	0.95	0.87							0.91	0.73	0.63	0.61	0.93	0.82	0.98	0.44	0.86	0.26

Table 3. Means of results obtained by participants on the benchmark test case (corresponding to harmonic means). The symmetric relaxed measure corresponds to the relaxed precision and recall measures of [4].

precision and recall reported in Table 3). The values are not anymore an average but a real precision and recall over all the tests. The numbers in the legend are the Mean Average Precision (MAP): the average precision for each correct retrieved correspondence. These new graphs represent well the effort made by the participants to keep a high precision in their results, and to authorise a loss of precision with a few correspondences with low confidence.

The results presented in Table 3 and those displayed in Figure 2 single out the same group of systems, ASMOV, RiMOM and AgrMaker, which seem to perform these tests at the highest level. Of these, ASMOV has slightly better results than the two others. So, this confirms the observations on raw results.

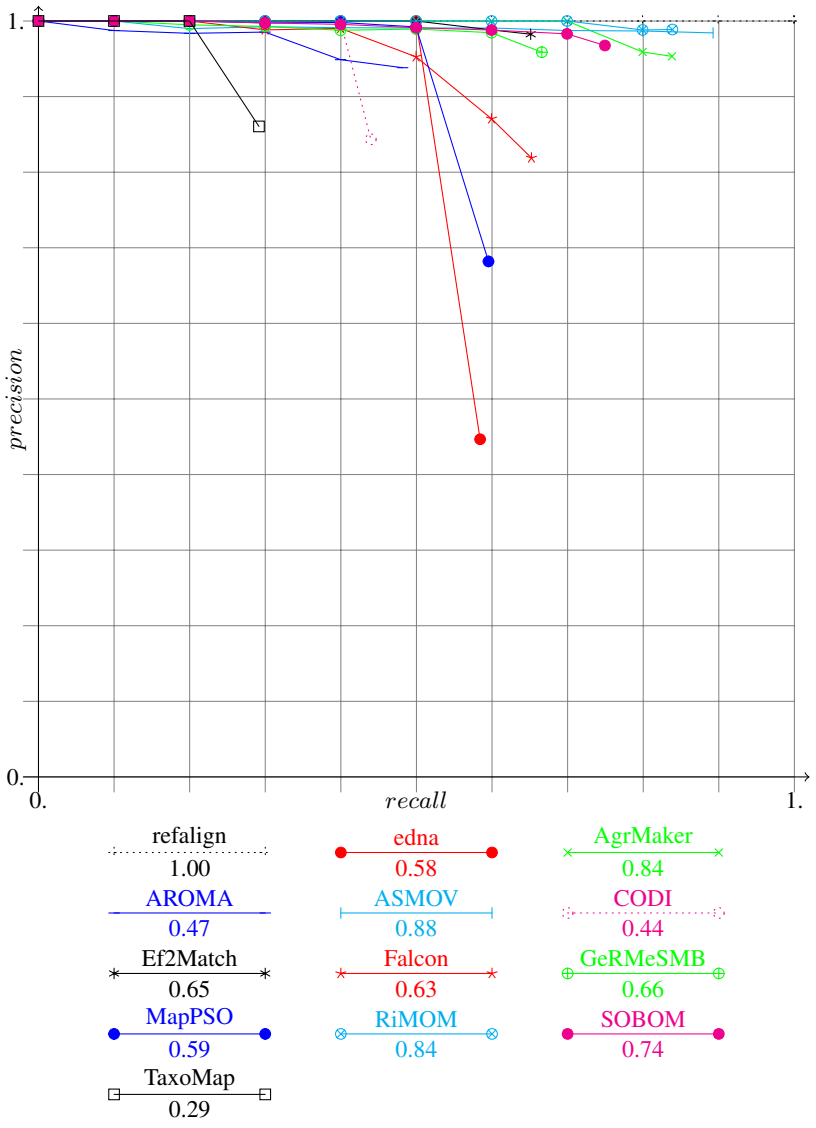

Fig. 2. Precision/recall graphs for benchmarks. The results given by the participants are cut under a threshold necessary for achieving $n\%$ recall and the corresponding precision is computed. Systems for which these graphs are not meaningful (because they did not provide graded confidence values) are drawn in dashed lines.

4 Anatomy

The anatomy track confronts existing matching technology with a specific type of ontologies from the biomedical domain. In this domain, a significant number of ontologies have been built covering different aspects of medical research.

4.1 Test data and experimental setting

The data set of this track has been used since 2007. For a detailed description we refer the reader to the OAEI 2007 [7] results paper. The ontologies of the anatomy track are the NCI Thesaurus describing the human anatomy, published by the National Cancer Institute (NCI), and the Adult Mouse Anatomical Dictionary, which has been developed as part of the Mouse Gene Expression Database project. Both resources are part of the Open Biomedical Ontologies (OBO). The alignment between these ontologies has been created by experts of the domain [2].

As in the previous years, we divided the matching task into four subtasks. Subtask #1 is obligatory for participants of the anatomy track, while subtask #2, #3 and #4 are again optional tasks.

Subtask #1 The matcher has to be applied with its standard settings.
Subtask #2 An alignment has to be generated that favors precision over recall.
Subtask #3 An alignment has to be generated that favors recall over precision.
Subtask #4 A partial reference alignment has to be used as additional input.

Notice that in 2010 we used the SEALS evaluation service for subtask #1. In the course of using the SEALS services, we published the complete reference alignment for the first time. In the future, we plan to include all subtasks in the SEALS modality. This requires to extend the interfaces of the SEALS evaluation service to allow for example an (incomplete) alignment as additional input parameter.

The harmonization of the ontologies applied in the process of generating a reference alignment (see [2] and [7]), resulted in a high number of rather trivial correspondences (61%). These correspondences can be found by very simple string comparison techniques. At the same time, we have a good share of non-trivial correspondences (39%). This is an important characteristic of the data set to be taken into account in the following analysis. The partial reference alignment used in subtask #4 is the union of all trivial correspondences and 54 non-trivial correspondences.

Due the experiences made in the past, we decided to slightly modify the test data set for the 2010 evaluation. We removed some doubtful subsumption correspondences and added a number of disjointness statement at the top of the hierarchies to increase the expressivity of the data set. Furthermore, we eliminated three incorrect correspondences. The reference alignment is now coherent with respect to the ontologies to be matched[7].

[7] We gratefully thank Elena Beisswanger (Jena University Language and Information Engineering Lab) for her thorough support on improving the quality of the data set. The modifications are documented at `http://webrum.uni-mannheim.de/math/lski/anatomy10/modifications2010.html`

4.2 Results

While the number of participants is nearly stable over four years, we find in 2010 more systems that participated for the first time (5 systems) than in the previous years (in average 2 systems). See Table 4 for an overview. Four of the newcomers participate also in other tracks, while NBJLM participates only in the Anatomy track. NBJLM is thus together with AgreementMaker (AgrMaker) a system that uses a track-specific parameter setting. Taking part in several tracks with a standard setting makes it obviously much harder to obtain good results in a specific track.

System	2007	2008	2009	2010
AFlood		√	√	
AgrMaker	√		+	+
AROMA		√	√	
AOAS	+			
ASMOV	√	√	√	√
BLOOMS				+
CODI				√
DSSim	√	√	√	
Ef2Match				+
Falcon AO	√			
GeRMeSMB				√
Kosimap			√	
Lily	√	√	√	
NBJLM				+
Prior+	√			
RiMOM	√	+	√	
SAMBO	+	+		
SOBOM			+	+
TaxoMap	√	√	√	+
X SOM	√			
Avg. F-measure	0.598	0.718	0.764	0.785

Table 4. Overview on anatomy participants from 2007 to 2010, a √-symbol indicates that the system participated, + indicates that the system achieved an F-measure ≥ 0.8 in subtask #1.

In the last row of Table 4, the average of F-measures per year in subtask #1 is shown. We observe significant improvements over time. However, the measured improvements decrease over time and seem to reach a top (2007 +12% → 2008 +5% → 2009 +2% → 2010). We have marked the participants with an F-measure ≥ 0.8 with a + symbol. Note that in each of the previous years, only two systems reached this level, while in 2010 six systems reached a higher value than 0.8.

Runtimes In the previous years, we reported about runtimes that have been measured by the participants. The differences we observed – from several minutes to several days – could not be explained by the use of different hardware. However, these differences

became less significant over the years and in 2009 all systems except one required between 2 and 30 minutes. Therefore, we abstained from an analysis of runtimes this year. In 2011, we plan to execute the matching systems on the SEALS platform to enable an exact measurement of runtimes not biased by differences in hardware equipment. So far we refer the reader interested in runtimes to the result papers of the participants.

Main results for subtask #1 The results for subtask #1 are presented in Table 5 ordered with respect to the achieved F-measure. In 2010, AgreementMaker (AgrMaker) generates the best alignment with respect to F-measure. Moreover, this result is based on a high recall compared to the systems on the following positions. This is a remarkable result, because even the SAMBO system of 2007 could not generate a higher recall with the use of UMLS. However, we have to mention again that AgreementMaker uses a specific setting for the anatomy track.

System	Task #1			Task #2			Task #3			Recall+	
	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F	#1	#3
AgrMaker*	0.903	0.853	0.877	0.962	0.751	0.843	0.771	0.874	0.819	0.630	0.700
Ef2Match	0.955	0.781	0.859	-	-	-	-	-	-	0.440	-
NBJLM*	0.920	0.803	0.858	-	-	-	-	-	-	0.569	-
SOBOM	0.949	0.778	0.855	-	-	-	-	-	-	0.433	-
BLOOMS	0.954	0.731	0.828	0.967	0.725	0.829	-	-	-	0.315	-
TaxoMap	0.924	0.743	0.824	0.956	0.689	0.801	0.833	0.774	0.802	0.336	0.414
ASMOV	0.799	0.772	0.785	0.865	0.757	0.808	0.717	0.792	0.753	0.470	0.538
CODI	0.968	0.651	0.779	0.964	0.662	0.785	0.782	0.695	0.736	0.182	0.383
GeRMeSMB	0.884	0.307	0.456	0.883	0.307	0.456	0.080	0.891	0.147	0.249	0.838

Table 5. Results for subtasks #1, #2 and #3 in terms of precision, recall (in addition recall+ for #1 and #3) and F-measure. Systems marked with a * do not participate in other tracks or have chosen a setting specific to this track. Note that ASMOV modified its standard settting in a very restricted way (activating UMLS as additional resource). Thus, we did not mark this system.

AgreementMaker is followed by three participants (Ef2Match, NBJLM and SOBOM) that share a very similar characteristic regarding F-measure and observed precision score. All of these systems clearly favor precision over recall. A further analysis has to clarify to which degree the alignments generated by these systems are overlapping as indicated by their precision/recall characteristics. Notice that these systems obtained better scores or scores that are similar to the results of the top systems in the previous years. One explanation can be seen in the fact that the organizers of the track made the reference alignment available to the participants. More precisely, participants could at any time compute precision and recall scores via the SEALS services to test different settings of their algorithms. On the one hand, this allows to improve a matching system by a constant formative evaluation in a direct feedback cycle, on the other hand, it might happen that a perfect configuration results in problems for different data sets.

Recall+ and further results In the following, we use again the recall+ measure as defined in [7]. It measures how many non trivial correct correspondences, not detectable by string equivalence, can be found in an alignment. The top three systems with respect to recall+ regarding subtask #1 are AgreementMaker, NBJLM and ASMOV. Only ASMOV has participated in several tracks with the same setting. Obviously, it is not easy to find a large amount of non-trivial correspondences with a standard setting.

In 2010, five system participated in subtask #3. The top three systems regarding recall+ in this task are GeRoMe-SMB (GeRMeSMB), AgreementMaker and ASMOV. Since a specific instruction about the balance between precision and recall is missing in the description of the task, the results vary to a large degree. GeRoMe-SMB detected 83.8% of the correspondences marked as non-trival, but at a precision of 8%. AgreementMaker and ASMOV modified their settings only slightly, however, they were still able to detect 70% and 53.8% of all non trivial correspondences.

In subtask #2, six systems participated. It is interesting to see that systems like ASMOV, BLOOMS and CODI generate alignments with slightly higher F-measure for this task compared to the submission for subtask #1. The results for subtask #2 for AgreementMaker are similar to the results submitted by other participants for subtask #1. This shows that many systems in 2010 focused on a similar strategy that exploits the specifics of the data set resulting in a high F-measure.

Only about half of the participants submitted results for subtask #2 and #3. This can be related to an unclear description of the expected results. In the future we have to think about an alternative description of the subtask together with a different kind of evaluation to increase participation.

Subtask #4 In the following, we refer to an alignment generated for subtask #n as A_n. In our evaluation we use again the method introduced in 2009. We compare both $A_1 \cup R_p$ and $A_4 \cup R_p$ with the reference alignment R.[8] Thus, we compare the situation where the partial reference alignment is added after the matching process against the situation where the partial reference alignment is available as additional resource exploited within the matching process. Note that a direct comparison of A_1 and A_4 would not take into account in how far the partial reference alignment was already included in A_1 resulting in a distorted interpretation.

System	Δ-Precision	Δ-Recall	Δ-F-measure
AgrMaker	$+0.025_{\ 0.904 \to 0.929}$	$-0.025_{\ 0.876 \to 0.851}$	$-0.002_{\ 0.890 \to 0.888}$
ASMOV	$+0.029_{\ 0.808 \to 0.837}$	$-0.016_{\ 0.824 \to 0.808}$	$+0.006_{\ 0.816 \to 0.822}$
CODI	$-0.002_{\ 0.970 \to 0.968}$	$+0.030_{\ 0.716 \to 0.746}$	$+0.019_{\ 0.824 \to 0.843}$
SAMBOdtf$_{2008}$	$+0.021_{\ 0.837 \to 0.856}$	$+0.003_{\ 0.867 \to 0.870}$	$+0.011_{\ 0.852 \to 0.863}$

Table 6. Changes in precision, recall and F-measure based on comparing $A_1 \cup R_p$ and A_4 against reference alignment R.

Results are presented in Table 6. Three systems participated in task #4 in 2010. Additionally, we added a row for the 2008 submission of SAMBOdtf. This system

[8] We use $A_4 \cup R_p$ – instead of using A_4 directly – to ensure that a system, which does not include the input alignment in the output, is not penalized.

had the best results measured in the last years. AgreementMaker and ASMOV use the input alignment to increase the precision of the final result. At the same time these systems filter out some correct correspondences, finally resulting in a slightly increased F-measure. This fits with the tendency we observed in the past years (compare with the results for SAMBOdtf in 2008). The effects of this strategy are not very strong. However, as argued in the previous years, the input alignment has a characterictis that makes it hard to exploit this information.

CODI has chosen a different strategy. While changes in precision are negligible, recall increases by 3%. Even though the overall effect is still not very strong, the system exploits the input alignment in the most effective way. However, the recall of CODI for subtask #1 is relatively low compared to the other systems. It is unclear whether the strategy of CODI would also work for the other systems where a ceiling effect might prevent the exploitation of the positive effects. We refer the interested reader to the results paper of the system for a description of the algorithm.

4.3 Conclusions

Overall, we see a clear improvement comparing this years results with the results of the previous years. This holds both for the "average participant" as well as for the top performer. A very positive outcome can be seen in the increased recall values. In addition to the evaluation experiments we reported, we computed the union of all submissions to subtask #1. For the resulting alignment we measured a precision of 69.7% and a recall of 92.7%. We added additionaly the correct correspondences generated in subtask #3 and reached a a recall of 97.1%. Combining the strategies used by different matching systems it is thus possible to detect nearly all correct correspondences.

The availability of the SEALS evaluation service surely had an effect on the results submitted in 2010. We have already argued about pros and cons. In the future, we plan to extend the data set of the anatomy track with additional ontologies and reference alignments to a more comprehensive and general track covering different types of biomedical ontologies. In particular, we will not publish the complete set of reference alignments to conduct a part of the evaluation experiment in the blind mode. This requires, however, to find and analyze interesting and well-suited data sets. The strategy to publish parts of the evaluation material and to keep other parts hidden seems to be the best approach.

5 Conference

The conference test set introduces matching several more-or-less expressive ontologies. Within this track the results of participants will be evaluated using diverse evaluation methods. At the time of writing this paper we have only completed two evaluation methods, i.e. classical evaluation with respect to a reference alignment, which have been made for the ontology pairs where this alignment is available and posterior manual evaluation for all ontology pairs using sampling across all matchers. Third we plan that the complete results will be submitted to a data mining tool for discovery of association hypotheses, taking into account specific matching patterns. Fourth, alignment incoherence will be analysed with the help of a logical reasoner.

5.1 Test data

The collection consists of sixteen ontologies in the domain of organizing conferences. Ontologies have been developed within the OntoFarm project[9]. In contrast to last year's conference data set, this year is supported by the SEALS evaluation service.

The main features of this test set are:

- *Generally understandable domain.* Most ontology engineers are familiar with organizing conferences. Therefore, they can create their own ontologies as well as evaluate the alignment among their 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 DL axioms of various kinds, which opens a way to use semantic matchers.

Ontologies differ in numbers of classes, of properties, in their logical expressivity, but also in underlying resources. Eleven 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 all correct correspondences (equivalence and/or subsumption) and/or "interesting correspondences" within a collection of ontologies describing the domain of organizing conferences.

This year, results of participants will be evaluated by four different methods of evaluation: evaluation based on a reference alignment, manual labeling, data mining method, and logical reasoning. Similarly to OAEI 2009, we have still 21 alignments (with some corrections in comparison with the previous year), which correspond to the complete alignment space between 7 ontologies from the data set. Manual evaluation will produce statistics such as precision and will also serve as input into evaluation based on data mining and will help in the process of improving and building a reference alignment. Results of participants will be checked with regard to their coherence. These evaluation methods are described at the data set result page.

5.2 Results

We had eight participants: AgreementMaker (AgrMaker), AROMA, ASMOV, CODI, Ef2Match, Falcon, GeRMeSMB and SOBOM. Here are some basic data, besides evaluations:

- All participants delivered all 120 alignments.
- CODI matcher delivered 'certain' correspondences, other matchers delivered correspondences with graded confidence values between 0 and 1

	t=0.2			t=0.5			t=0.7		
	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.
AgrMaker	52%	68%	57%	52%	68%	57%	60%	60%	59%
AROMA	37%	50%	41%	38%	5%	42%	40%	20%	25%
ASMOV	55%	68%	60%	23%	7%	1%	28%	4%	6%
CODI	88%	52%	64%	88%	52%	64%	88%	52%	64%
Ef2Match	52%	66%	57%	52%	66%	57%	52%	66%	57%
Falcon	62%	61%	61%	62%	61%	61%	62%	61%	61%
GeRMeSMB	36%	53%	42%	36%	53%	42%	37%	53%	43%
SOBOM	34%	65%	44%	67%	17%	26%	0%	0%	0%

Table 7. Recall, precision and F-measure for three different confidence thresholds.

Evaluation based on reference alignment We evaluated the results of participants against a reference alignment. It includes all pairwise combinations of different 7 ontologies (i.e. 21 alignments).

In Table 7, there are traditional precision, recall, and F-measure computed for three different thresholds of confidence values (0.2, 0.5, and 0.7).[10]

matcher	confidence threshold	Prec.	Rec.	FMeas.
AgrMaker	0.61	53%	68%	58%
AROMA	0.45	37%	50%	42%
ASMOV	0.17	53%	71%	60%
CODI	*	88%	52%	64%
Ef2Match	0.83	63%	61%	61%
Falcon	0.92	80%	52%	62%
GeRMeSMB	0.77	39%	53%	44%
SOBOM	0.37	60%	56%	57%

Table 8. Confidence threshold, precision and recall for optimal F-measure for each matcher.

For a better comparison, we established the confidence threshold which provides the highest average F-measure (Table 8). Precision, Recall, and F-measure are given for this optimal confidence threshold. The dependency of F-measure on confidence threshold can be seen from Figure 3. There is one asterisk in the column of confidence threshold for matcher CODI which did not provide graded confidence.

In conclusion, the matcher with the highest average F-measure (62%) is the CODI which did not provide graded confidence values. Other matchers are very close to this score (e.g. Falcon with 62% of F-Measure, Ef2Match with 61% of F-Measure, ASMOV with 60% of F-Measure). However, we should take into account that this evaluation has been made over a subset of all alignments (one fifth).

Comparison with previous years We can compare performance of participants wrt. last two years (2008, 2009). There are three matchers which also participated in last two

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

[10] Alignments which are empty due to thresholding are considered as having zero precision.

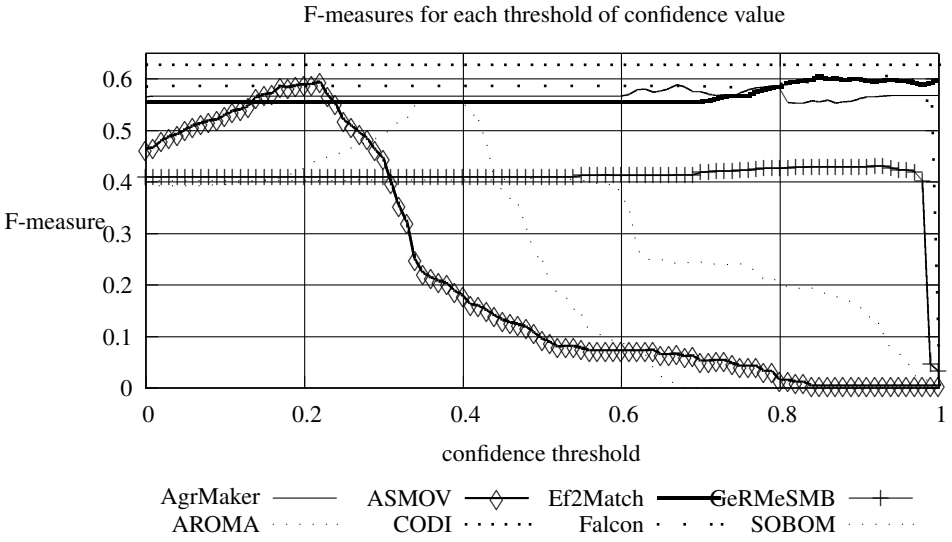

Fig. 3. F-measures depending on confidence.

years. ASMOV participated in all three consecutive years with increasing highest average F-measure: from 43% in 2008 and 47% in 2009 to 60% in 2010. AgreementMaker participated with 57% in 2009 and 58% in 2010 regarding highest average F-measure. Finally, AROMA participated with the same highest average F-measure in both years, 2009 and 2010.

Evaluation based on posterior manual labeling This year we take the most secure, i.e., with highest confidence, correct correspondences as a population for each matcher. Particularly, we evaluate 100 correspondences per matcher randomly chosen from all correspondences of all 120 alignments with confidence 1.0 (sampling). Because AROMA, ASMOV, Falcon, GeRMeSMB and SOBOM do not have enough correspondences with 1.0 confidence we take 100 correspondences with highest confidence. For all of these matchers (except ASMOV where we found exactly 100 correspondences with highest confidence values) we sampled over their population.

In table 9 you can see approximated precisions for each matcher over its population of best correspondences. N is a population of all the best correspondences for one matcher. n is a number of randomly chosen correspondences so it is 100 best correspondences for each matcher. TP is a number of correct correspondences from the sample, and P* is an approximation of precision for the correspondences in each population; additionally there is a margin of error computed as: $\frac{\sqrt{(N/n)-1}}{\sqrt{N}}$ based on [15].

From Table 9 we can conclude that CODI, Falcon and AgreementMaker have the best precision (higher than 90%) over their 100 more confident correspondences.

matcher	AgrMaker	AROMA	ASMOV	CODI	Ef2Match	Falcon	GeRMeSMB	SOBOM
N	804	108	100	783	1236	127	110	105
n	100	100	100	100	100	100	100	100
TP	92	68	86	98	79	96	30	82
P*	92% ±9.4%	68% ±2.7%	86%	98% ±9.3%	79% ±9.6%	96% ±4.6%	30% ±3.0%	82% ±2.2%

Table 9. Approximated precision for 100 best correspondences for each matcher.

6 Directory

The directory test case aims at providing a challenging task for ontology matchers in the domain of large directories to show whether ontology matching tools can effectively be applied for the integration of "shallow ontologies". The focus of this task is to evaluate performance of existing matching tools in real world taxonomy integration scenario.

6.1 Test set

As in previous years [8; 7; 3; 6], the data set exploited in the directory matching task was constructed from Google, Yahoo and Looksmart web directories following the methodology described in [10]. 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. This year, however, we have used two modalities:

1. Small task: this modality corresponds to the last years directory tracks and aims at testing multiple specific node matching tasks.
2. Single task: this modality contains only one matching task.

Both modalities present the following common characteristics:

– Simple relationships. Basically web directories contain only one type of relationship so called "classification relation".
– Vague terminology and modeling principles: The matching tasks incorporate the typical "real world" modeling and terminological errors.

Small task modality 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 [10], human annotators consider only 2265 correspondences instead of the full matching problem.

The specific characteristics of the data set for the small task modality 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 the equivalence relation for all the matching tasks.

Single task modality These directories correspond to a superset of all the "small" directories contained in the small tasks modality. The aim of this modality is to test the ability of current matching systems to handle and match big directories. This modality confronts the participating systems with a realistic scenario that can be found in many commercial application areas, involving web directories.

The specific characteristics of the data set for the single task modality are:

- A single matching task where the aim is to find the correspondences between the directory nodes, where each directory contains 2854 and 6555 nodes respectively.
- Reference correspondences for the matching task. This task includes, besides the equivalence relation, more general and less general relations.

6.2 Results

Small tasks modality In OAEI-2010, 3 out of 15 matching systems participated on the web directories test case, while in OAEI-2009 7 out of 16, in OAEI-2008, 7 out of 13, in OAEI-2007, 9 out of 17, 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 4. These indicators have been computed following the TaxMe2 [10] methodology, with the help of the Alignment API [5], version 3.4.

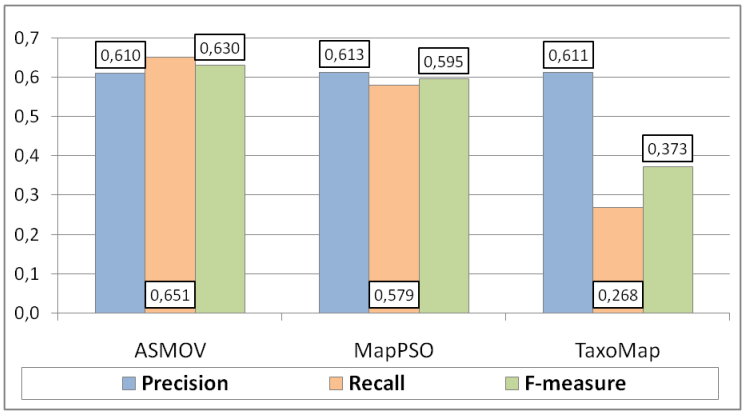

Fig. 4. Matching quality results.

We can observe from Table 10, that ASMOV has maintained its recall, but increased its precision by 1 point in comparison to 2009. MapPSO has increased its recall (+27) and precision (+7) values, resulting in a 20 points increase in the F-measure from its last participation in 2008. TaxoMap has decreased its recall (-7) but increased its precision (+3), resulting in an overall decrease of F-measure (-6) from its last participation in 2009. ASMOV is the system with the highest F-measure value in 2010.

Table 10 shows that in total 24 matching systems have participated during the 6 years (2005 - 2010) of the OAEI campaign in the directory track. In total, 40 submissions from different systems have been received over the past 6 years. No single system has participated in all campaigns involving the web directory dataset (2005 - 2010). A

total of 15 systems have participated only one time in the evaluation, 5 systems have participated 3 times (DSSIM, Falcon, Lily, RiMOM and TaxoMap), and only 1 system has participated 4 times (ASMOV).

System	**Recall**						**Precision**					**F-Measure**				
Year →	2005	2006	2007	2008	2009	2010	2006	2007	2008	2009	2010	2006	2007	2008	2009	2010
aflood					0.40					0.57					0.47	
ASMOV			0.44	0.12	0.65	0.65		0.59	0.64	0.60	0.61		0.50	0.20	0.63	0.63
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	0.41			0.60	0.60	0.60			0.41	0.49	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				
kosimap					0.52					0.62					0.56	
Lily			0.54	0.37	0.33			0.57	0.59	0.57			0.55	0.46	0.42	
MapPSO				0.31		0.58			0.57		0.61			0.40		0.60
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		
SOBOM					0.42					0.59					0.49	
TaxoMap				0.34	0.34	0.27			0.59	0.59	0.61			0.43	0.43	0.37
X-SOM			0.29					0.62					0.39			
Average	*0.22*	*0.26*	*0.50*	*0.30*	*0.44*	*0.50*	*0.35*	*0.57*	*0.59*	*0.59*	*0.61*	*0.29*	*0.49*	*0.39*	*0.50*	*0.53*
#	*7*	*7*	*9*	*7*	*7*	*3*	*7*	*9*	*7*	*7*	*3*	*7*	*9*	*7*	*7*	*3*

Table 10. 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 5 and Table 10, this year there is a small increase (2%) in the average precision, in comparison to 2007 and 2008. The average recall in 2010 increased in comparison to 2009, reaching the same highest average recall value as in 2007. Considering F-measure, results for 2009 show the highest average in the 5 years (2006 to 2010). Notice that in 2005 the data set allowed only the estimation of recall, therefore Figure 5 and Table 10 do not contain values of precision and F-measure for 2005.

A comparison of the results from 2006 - 2010 for the top-3 systems of each year based on the highest values of the F-measure indicator is shown in Figure 6. An important note is that since there are only 3 participants this year, they all made their ways into the top three. The comparison of the top three participants has being made since 2006, therefore we keep the same comparison (and not the top 2, for example) for historical

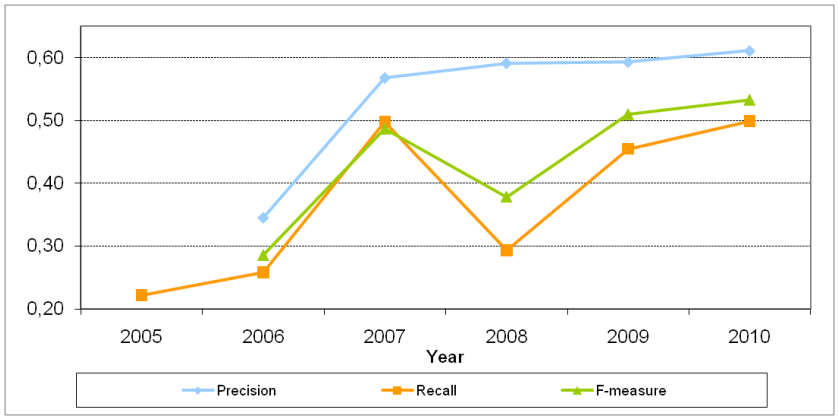

Fig. 5. Average results of the participating systems per year.

reasons. The quality of the best F-measure result of 2010 (0.63) achieved by ASMOV is equal to the best F-measure of 2009 by the same system, higher than the best F-measure of 2007 by DSSim (0.49) and than that of 2006 by Falcon (0.43), but still lower than the best F-measure of 2007 (0.71) by OLA_2. All three participating systems have achieved the same precision in 2010 (0.61), but this precision is lower than the best values of 2009 (0.62) by kosimap, in 2008 (0.64) by ASMOV and in 2007 by both OLA_2 and X-SOM. Finally, for what concerns recall, the best result of 2010 achieved by ASMOV (0.65) is equal to the best value of 2009 (0.65) also achieved by ASMOV, higher than the best value of 2008 (0.41) demonstrated by DSSim and the best value in 2006 (0.45) by Falcon, but still lower than the best result obtained in 2007 (0.84) obtained by OLA_2.

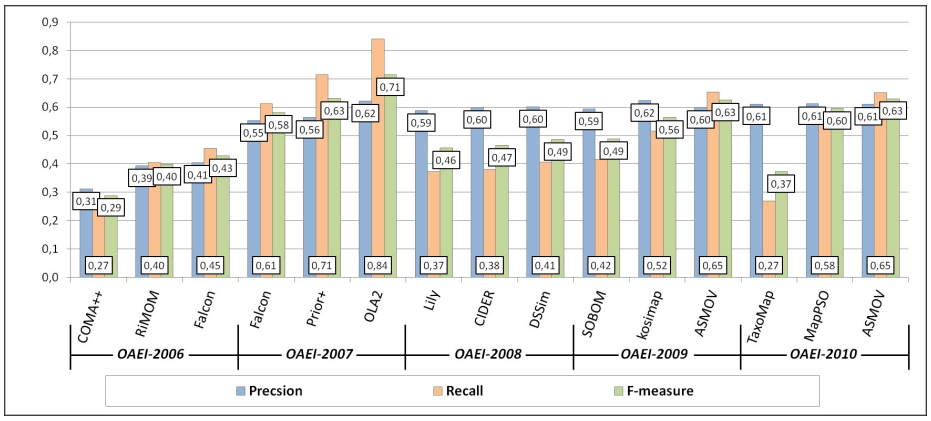

Fig. 6. Comparison of matching quality results in 2006 - 2009.

Figure 7 shows the yearly averages of the top 3 systems where we can see that the best values for recall and F-measure were obtained in 2007. The precision value made a significant increase also in 2007 in comparison to the value of 2006, but since 2007 only small steady increases where achieved each year.

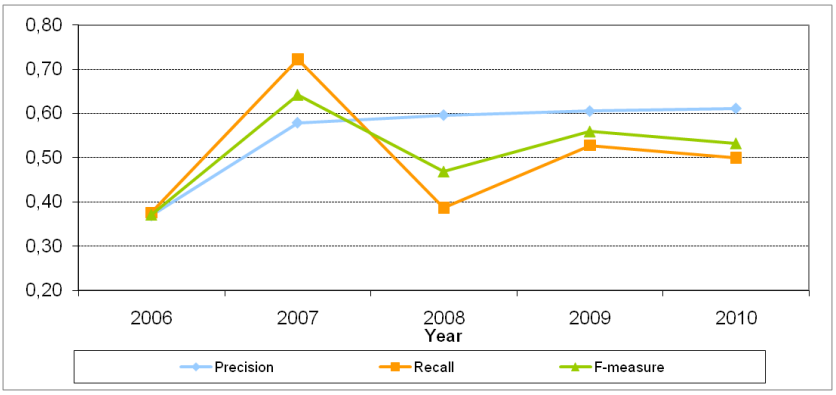

Fig. 7. Average results of the top-3 systems per year.

Partitions of positive and negative correspondences, according to the system results, are presented in Figure 8 and Figure 9, respectively. Figure 8 shows that the systems managed to discover only 67% of the total number of positive correspondences (Nobody = 33%). Only 27% of positive correspondences were found by all three participating systems. The percentage of positive correspondences found by the systems this year is slightly lower than the values of 2009, when 68% of the positive correspondences where found [6], but still higher than the values of 2008, when 54% of the positive correspondences where found [3]. Figure 9 shows that more than half (59%) of the negatives correspondences were not found by the systems (correctly) in comparison to 56% not found in 2009). Figure 9 also shows that all participating systems found 16% of the negative correspondences, i.e., mistakenly returned them as positive, in comparison to 17% in 2009. These two observations explain the small increase in precision in Figure 5 and Figure 7. The last two observations also suggest that the discrimination ability of the dataset remains still high as in previous years.

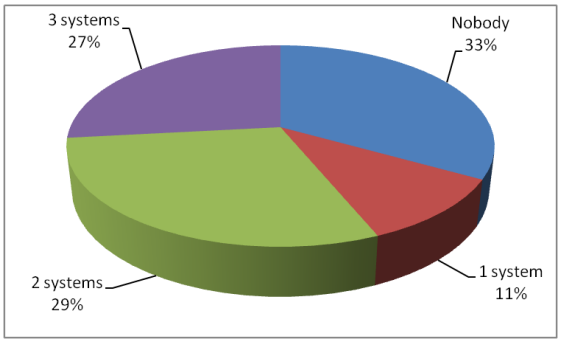

Fig. 8. Partition of the system results on positive correspondences.

Let us now compare partitions of the system results in 2006 - 2010 on positive and negative correspondences, as shown in Figure 10 and Figure 11, respectively.

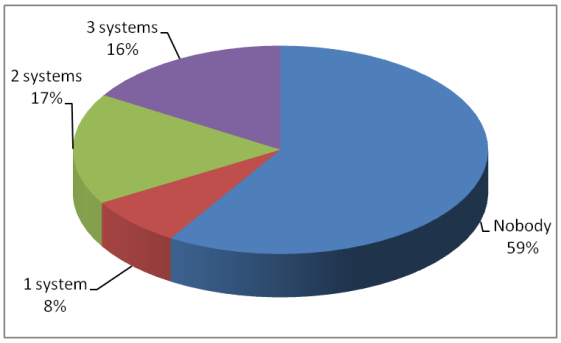

Fig. 9. Partition of the system results on negative correspondences.

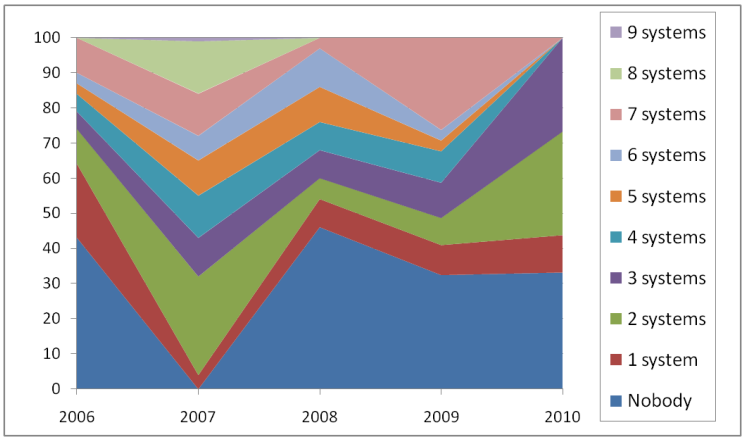

Fig. 10. Comparison of partitions of the system results on positive correspondences in 2006 - 2009.

Figure 10 shows that 33% of positive correspondences have not been found by any of the matching systems this year. This value is better that the values of 2006 (43%) and 2008 (46%) but worse than 2009 (32%). In 2007, all the positive correspondences have been collectively found; these results (2007) were exceptional because the participating systems altogether had a full coverage of the expected results and very high precision and recall. Unfortunately, the best systems of 2007 did not participate this year (nor in 2008 and 2009) and the other systems do not seem to cope with the results of 2007.

Figure 11 shows that this year 59% of the negatives correspondences were correctly not found. There is an increase in comparison to the value of 2009 (56%) but a decrease in comparison to the value of 2008, when 66% of the negatives correspondences where not found, being the best value in all years (2006 to 2010). This year 16% of the negative correspondences were mistakenly found by all the (3) participating systems, being the best value that of 2008 (1% for all (7) participating systems). An interpretation of these observations could be that the set of participating systems in 2010 seems to have found a good balance between being “cautious” (not finding negatives) and be-

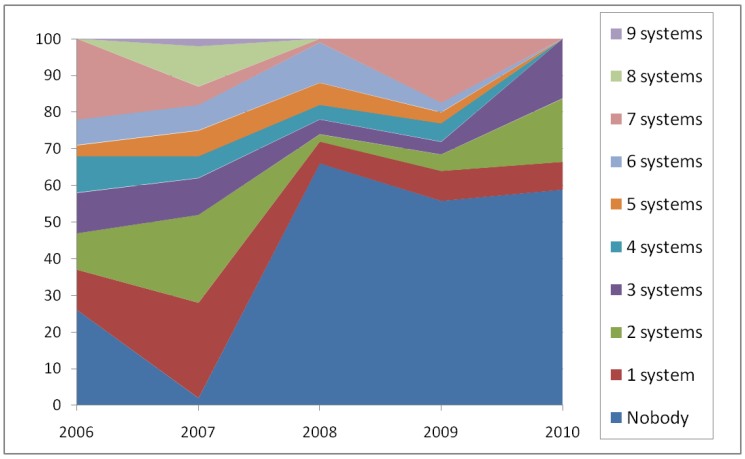

Fig. 11. Comparison of partitions of the system results on negative correspondences in 2006 - 2010.

ing "brave" (finding positives), resulting in average increases on precision, recall and F-measure as shown in Figure 5. In average, in 2010 the participants have a more "cautious" strategy of all years except 2008, being a little bit more "brave" than in 2007 and 2008. In 2007, we can observe that the set systems showed the most "brave" strategy in discovering correspondences of all the yearly evaluation initiatives, when the set of positive correspondences was fully covered, but covering mistakenly also 98% of the negative correspondences.

Single task modality [Work in progress (TBA soon)].

6.3 Comments

This year the average performance of the participants on the small tasks (given by the increase in precision and F-measure in Figure 5) is the best of all 5 years (2006 to 2010). This suggests that the set of participating systems has found a balance between a "brave and cautious" behavior for discovering correspondences. However, the value for the F-measure (0.53) indicates that there is still room for further improvements. In comparison to 2009, there is an increase of 2% in F-measure where the average F-measure was (0.51). Finally, as partitions of positive and negative correspondences indicate (see Figure 8 and Figure 9), the dataset still retains a good discrimination ability, i.e., different sets of correspondences are still hard for the different systems.

7 Instance matching

The instance matching track was included into the OAEI campaigns for the second time. The goal of the track is to evaluate the performance of different tools on the task of matching RDF individuals which originate from different sources but describe the

same real-world entity. With the development of the Linked Data initiative, the growing amount of semantic data published on the Web and the need to discover identity links between instances from different repositories, this problem particularly gained importance in the recent years. Unlike in the other tracks, the instance matching tests specifically focus on ontology ABox. However, the problems which have to be resolved in order to match instances correctly can originate at the schema level (use of different properties and classification schemas) as well as at the data level (e.g., different format of values). This year, the track includes two subtracks. The first subtrack (data interlinking - DI) aims at testing the performance of tools on large-scale real-world datasets published according to the Linked Data principles. The second one (IIMB & PR) represents a set of artificially generated and real test cases respectively. They are designed to illustrate all common cases of discrepancies between individual descriptions (different value formats, modified properties, different classification schemas). The list of participants to the Instance Matching track is shown in Table 11.

System	DI	IIMB_SMALL	IIMB_LARGE	PR
ASMOV		√	√	√
ASMOV_D				√
CODI		√	√	√
LN2R				√
ObjectCoref	√			√
RiMOM	√	√	√	√

Table 11. Participants in the instance matching track.

7.1 Data interlinking track (DI)

Data interlinking is known under many names according to various research communities: equivalence mining, record linkage, object consolidation and corefernece resolution to mention the most used ones. In each case, these terms are used for the task of finding equivalent entities in or across datasets. As the quantity of datasets published on the Web of data dramatically increases, the need for tools helping to interlink resources become bigger. It is becoming particularly important to maximize the automation of the interlinking process in order to be able to follow this expansion.

For the second year, OAEI proposes a data interlinking track dedicated to interlink datasets published on the Web of data. This year, we propose to interlink four datasets together. We have selected datasets for their potential to be interlinked, for the availability of curated interlinks between them, and for their size. All datasets are on the health-care domain and all of them contain information about drugs. Below is a more detailed presentation of the datasets (See [12] for more details on the datasets).

dailymed is published by the US National Library of Medecine and contains information about marketed drugs. Dailymed contains information on the chemical structure, mechanism of action, indication, usage, contraindications and adverse reactions for the drugs.

diseasome contains information about 4300 disorders and genes.

drugbank a repository of more than 5000 drugs approved by the US Federal Drugs Agency. It contains information about chemical, pharmaceutical and pharmacological data along with the drugs data.

sider was originally published on flat files before being converted as linked-data through a relational database. It contains information on marketed drugs and their recorded adverse reactions.

These datasets were semi-automatically interlinked using the tools Silk [16] and ODD Linker [11] providing the reference alignments for this track and participants were asked to retrieve these links using an automatic method.

Only two systems participated the data interlinking track, probably due to the difficulties of matching large collections of data: ObjectCoref and RiMOM. The results of these systems are shown in Figure 12.

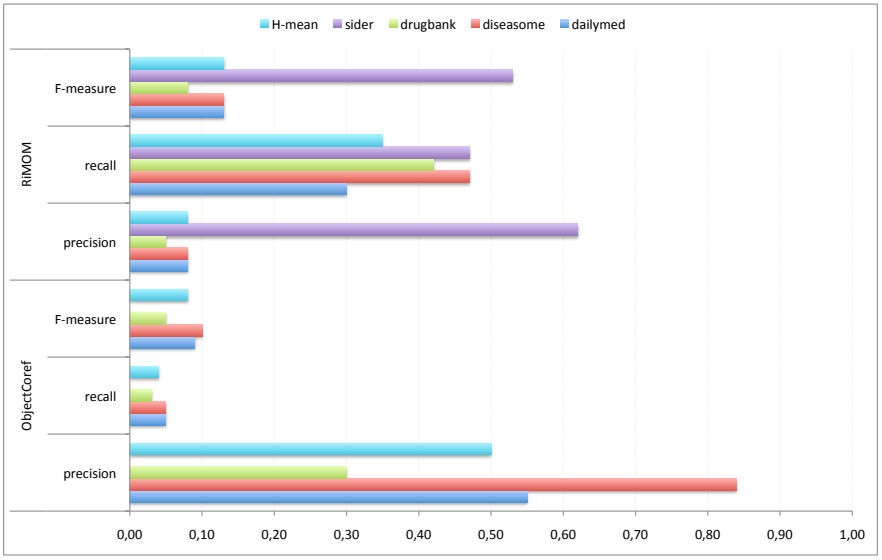

Fig. 12. Results of the DI subtrack.

The results are very different for the two systems, with ObjectCoref being better in precision and RiMOM being better in recall. In general, anyway, we have quite bad results for both the systems. A difficult task with real interlinked data is to understand if the results are bad because of a weakness of the matching system or because links can be not very reliable. In any case, what we can conclude from this experience with linked data is that a lot of work is still required in three directions: i) providing a reliable mechanism for systems evaluation; ii) improving the performances of matching systems in terms of both precision and recall; iii) work on the scalability of matching techniques in order to make affordable the task of matching large collections of real data. Starting from these challenges, data interlinking will be one one the most important future directions for the instance matching evaluation initiative.

7.2 OWL data track (IIMB & PR)

The OWL data track is focused on two main goals:

1. to provide an evaluation dataset for various kinds of data trasformations, including value trasformations, structural tranformations and logical transformations;
2. to cover a wide spectrum of possible techniques and tools.

To this end, we provided two groups of datasets, the ISLab Instance Matching Benchmark (IIMB) and the Person-Restaurants benchmark (PR). In both cases, participants were requested to find the correct correspondences among individuals of the first knowledge base and individuals of the other. An important task here is that some of the transformations require automatic reasoning for finding the expected alignments.

IIMB. IIMB is composed of a set of test cases, each one represented by a set of instances (i.e., an OWL ABox) built from an initial dataset of real linked data extracted from the web. Then, the ABox is automatically modified in several ways by generating a set of new ABoxes, called *test cases*. Each test case is produced by transforming the individual descriptions in the reference ABox in new individual descriptions that are inserted in the test case at hand. The goal of transforming the original individuals is twofold: on one side, we provide a simulated situation where data referred to the same objects are provided in different data sources; on the other side, we generate a number of datasets with a variable level of data quality and complexity. IIMB provides transformation techniques supporting the modifications of data property values, the modification of number and type of properties used for the individual description, and the modification of the individuals classification. The first kind of transformations is called *data value transformation* and it aims at simulating the fact that data depicting the same real object in different data sources may be different because of data errors or because of the usage of different conventional patterns for data representation. The second kind of transformation is called *data structure transformation* and it aims at simulating the fact that the same real object may be described using different properties/attributes in different data sources. Finally, the third kind of transformation, called *data semantic transformation*, simulates the fact that the same real object may be classified in different ways in different data sources.

The 2010 edition of IIMB is a collection of OWL ontologies consisting of 29 concepts, 20 object properties, 12 data properties and thousands of individuals divided into 80 test cases. In fact, in IIMB 2010, we have defined 80 test cases, divided into 4 sets of 20 test cases each. The first three sets are different implementations of data value, data structure and data semantic transformations, respectively, while the fourth set is obtained by combining together the three kinds of transformations. IIMB 2010 is created by extracting data from Freebase, an open knowledge base that contains information about 11 Million real objects including movies, books, TV shows, celebrities, locations, companies and more. Data extraction has been performed using the query language JSON together with the Freebase JAVA API[11]. The benchmark has been generated in a small version consisting in 363 individuals and in a large version containing

[11] http://code.google.com/p/freebase-java/

1416 individuals. In Figures 13 and 14 we report the results over the large version that are quite similar to the small one.

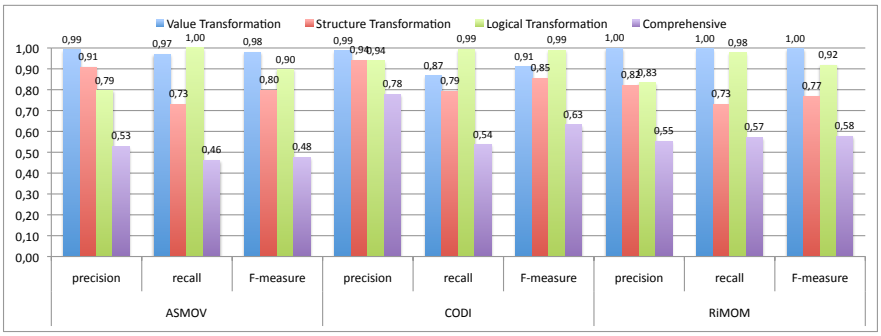

Fig. 13. Results of the IIMB subtrack.

The participation in IIMB was limited to ASMOV, CODI and RiMOM systems. All the systems obtained very good results when dealing with data value transformations and logical transformations, both in terms of precision and in terms of recall. Instead, in case of structural transformations (e.g., property value deletion of addition, property hierarchy modification) and of the combination of different kinds of transformations we have worst results, especially concerning recall. Looking at the results, it seems that the combination of different kinds of heterogeneity in data descriptions is still an open problem for instance matching systems. The three matching systems seems comparable in terms of quality of results.

PR. The Person-Restaurants benchmark is composed of three subsets of data. Two datasets (Person 1 and Person 2) contain personal data. The Person 1 dataset is created with the help of the Febrl project example datasets[12]. It contains original records of people and modified duplicate records of the same entries. The duplicate record set contains one duplicate per original record, with a maximum of one modification per duplicate record and a maximum of one modification per attribute. Person 2 is created as Person 1, but this time we have a maximum of 3 modifications per attribute, and a maximum of 10 modifications per record. The third dataset (Restaurant) is created with the help of 864 restaurant records from two different data sources (Fodor and Zagat restaurant guides)[13]. Restaurants are described by name, street, city, phone and restaurant category. Among these, 112 record pairs refer to the same entity, but usually display certain differences. In all the datasets the number of records is quite limited (about 500/600 entries). Results of the evaluation are shown in Figure 15.

The PR subtrack of the instance matching task was quite successful in terms of participation, in that all the five systems sent their results for this subtrack[14]. This is due

[12] Downloaded from http://sourceforge.net/projects/febrl/

[13] They can be downloaded from http://userweb.cs.utexas.edu/users/ml/riddle/data.html

[14] ASMOV sent a second set of results referred as ASMOV_D. They are the same as ASMOV but alignments are generated using the descriptions available in the TBOX

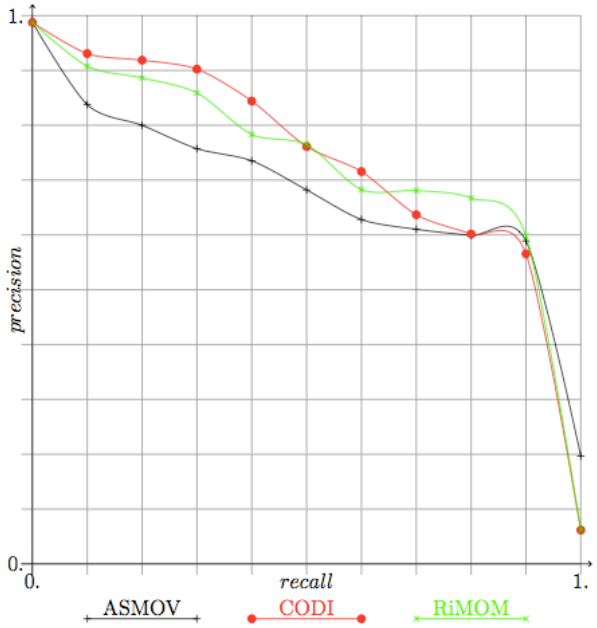

Fig. 14. Precision/recall of tools participating in the IIMB subtrack.

also to the fact that the PR datasets contain a small number of instances to be matched, resulting in a matching task that is affordable in terms of time required for comparisons. The results are good for all the systems with best performances obtained by RiMOM followed by ObjectCoref and LN2R. ASMOV and CODI instead have quite low values of F-measure in case of the Person 2 dataset. This is mainly due to low performances in terms of recall. These low values of recall depend on the fact that in Person 2 more than one matching counterpart was expected for each person record in the reference dataset.

8 Lesson learned and suggestions

We have seriously implemented the promises of last year with the provision of the first automated tool for evaluating ontology matching, the SEALS evaluation service, which have been used for three different data sets. We will continue on this path. We also took into account two other lessons: having rules for submitting data sets and rules for declaring them unfruitful that are published on OAEI web site. There still remain one lesson not really taken into account that we identify with an asterisk (*) and that we will tackle next year.

The main lessons from this year are:

A) We were not sure that switching to an automated evaluation would preserve the success of OAEI, given that the effort of implementing a web service interface was required from participants. This has been the case.

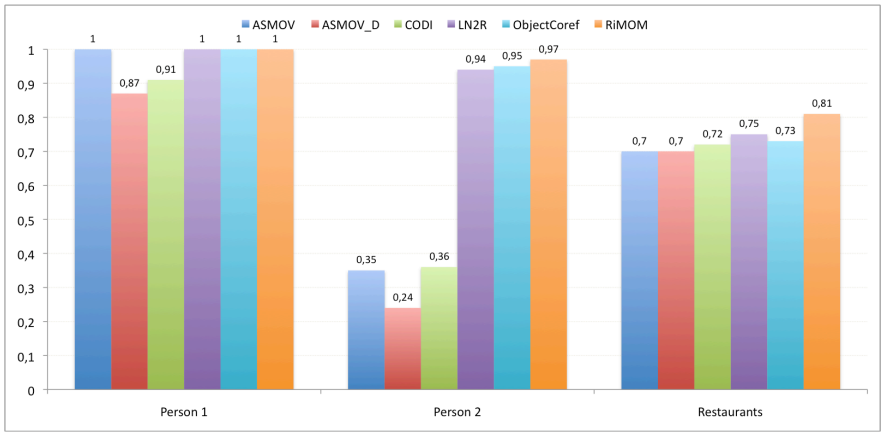

Fig. 15. Results of tools participating in the PR subtrack in terms of F–measure.

B) The SEALS service render easier the evaluation execution on a short period because participants can improve their systems and get results in real time. This is to some degree also possible for a blind evaluation. This is very valuable.

C) The trend that there are more matching systems able to enter such an evaluation seems to slow down. There have been not many new systems this year but on specialized topics. There can be two explanations: the field is shrinking or the entry ticket is too high.

D) We still can confirm that systems that enter the campaign for several times tend to improve over years. But we can also remark that they continue to improve (on data sets in which there still is a progress margin).

*E) The benchmark test case is not discriminant enough between systems. Next year, we plan to introduce controlled automatic test generation in the SEALS evaluation service and think that this will improve the situation.

F) SEALS participants were invited to register the information about their tools in the SEALS portal. However, some developers had registered their tool information but have not used the SEALS evaluation service neither for testing their tools nor for registering their final results. We contacted these developers, who had answered that they did not have enough time for preparing their tools. Again, the effort of implementing the web service interface and fixing all networks problems for making the service available could be one of the reasons why these developers have registered for participating in the campaign, but finally they did not do.

G) Not all systems followed the general rule to use the same set of parameters in all tracks. In addition, there are systems participating only in one track for which they are specialized. A fair comparison of general-purpose systems, specialized systems and optimally configured systems might require to rethink the application of this rule.

9 Future plans

There are several plans for improving OAEI. The first ones are related to the development of the SEALS services. In the current setting, runtime and memory consumption cannot be correctly measured because a controlled execution environment is missing. Further versions of the SEALS evaluation service will include the deployment of tools in such a controlled environment. As initially planned for last year, we plan to supplement the benchmark test with an automatically generated benchmark that would provide more challenge for participants. We also plan to generalize the use of the platform to other data sets.

In addition, we would like to have again a data set for evaluating tasks which requires alignments containing other relations than equivalence.

10 Conclusions

Confirming the trend of previous years, the number of systems, and tracks they enter in, seems to stabilize. 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).

The trend of number of tracks entered by participants went down again: 2.6 against 3.25 in 2009, 3.84 in 2008 and 2.94 in 2007. This figure of around 3 out of 8 may be the result of either the specialization of systems It is not the result of the short time allowed to the campaign, since the SEALS evaluation service has more run than what the participants registered.

All participants have provided a 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 also warmly thank Laura Hollinck, Véronique Malaisé and Willem van Hage for preparing the vlcr test case which has been cancelled.

We are grateful to Martin Ringwald and Terry Hayamizu for providing the reference alignment for the anatomy ontologies and thank Elena Beisswanger (Jena University

Language and Information Engineering Lab) for her thorough support on improving the quality of the data set.

We are grateful to Dominique Ritze (University of Mannheim) for participating in extension of reference alignment for the conference track.

We thank Andriy Nikolov and Jan Noessner for providing data in the process of constructing the IIMB dataset and we thank Heiko Stoermer and Nachiket Vaidya for providing the PR dataset for Instance Matching.

We also thank the other members of the Ontology Alignment Evaluation Initiative Steering committee: Wayne Bethea (John Hopkins University, USA), Lewis Hart (AT&T, USA), Tadashi Hoshiai (Fujitsu, Japan), Todd Hughes (DARPA, USA), Yannis Kalfoglou (Ricoh laboratories, UK), John Li (Teknowledge, USA), Miklos Nagy (The Open University (UK), Natasha Noy (Stanford University, USA), Yuzhong Qu (Southeast University (China), York Sure (Leibniz Gemeinschaft, Germany), Jie Tang (Tsinghua University (China), Raphaël Troncy (Eurecom, France), and Petko Valtchev (Université du Québec Montréal, Canada). George Vouros (University of the Aegean, Greece).

Jérôme Euzenat, Christian Meilicke, Heiner Stuckenschmidt and Cássia Trojahn dos Santos have been partially supported by the SEALS (IST-2009-238975) European project.

Ondřej Šváb-Zamazal and Vojtěch Svátek were supported by the IGA VSE grant no.20/08 “Evaluation and matching ontologies via patterns”.

References

1. Ben Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Proceedings of the K-Cap Workshop on Integrating Ontologies*, Banff (CA), 2005.
2. Oliver Bodenreider, Terry Hayamizu, Martin Ringwald, Sherri De Coronado, and Songmao Zhang. Of mice and men: Aligning mouse and human anatomies. In *Proc. American Medical Informatics Association (AIMA) Annual Symposium*, pages 61–65, 2005.
3. Caterina Caracciolo, Jérôme Euzenat, Laura Hollink, Ryutaro Ichise, Antoine Isaac, Véronique Malaisé, Christian Meilicke, Juan Pane, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, and Vojtech Svátek. Results of the ontology alignment evaluation initiative 2008. In *Proc. 3rd International Workshop on Ontology Matching (OM-2008), collocated with ISWC-2008*, Karlsruhe (Germany), 2008.
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, Alfio Ferrara, Laura Hollink, Antoine Isaac, Cliff Joslyn, Véronique Malaisé, Christian Meilicke, Andriy Nikolov, Juan Pane, Marta Sabou, François Scharffe, Pavel Shvaiko, Vassilis Spiliopoulos, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, Cássia Trojahn dos Santos, George Vouros, and Shenghui Wang. Results of the ontology alignment evaluation initiative 2009. In Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Natasha Noy, and Arnon Rosenthal, editors, *Proc. 4th ISWC workshop on ontology matching (OM), Chantilly (VA US)*, pages 73–126, 2009.

7. 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 *Proc. 2nd International Workshop on Ontology Matching (OM-2008), collocated with ISWC-2007*, pages 96–132, Busan (Korea), 2007.
8. 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 *Proc. 1st International Workshop on Ontology Matching (OM-2006), collocated with ISWC-2006*, pages 73–95, Athens, Georgia (USA), 2006.
9. Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer, Heidelberg (DE), 2007.
10. 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):137–157, 2009.
11. Oktie Hassanzadeh, Reynold Xin, Renée J. Miller, Anastasios Kementsietsidis, Lipyeow Lim, and Min Wang. Linkage query writer. *PVLDB*, 2(2):1590–1593, 2009.
12. Anja Jentzsch, Jun Zhao, Oktie Hassanzadeh, Kei-Hoi Cheung, Matthias Samwald, and Bo Andersson. Linking open drug data. In *Proceedings of Linking Open Data Triplification Challenge at the I-Semantics 2009*, 09 2009.
13. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the ISWC Workshop on Evaluation of Ontology-based Tools (EON)*, Hiroshima (JP), 2004.
14. Cássia Trojahn dos Santos, Christian Meilicke, Jérôme Euzenat, and Heiner Stuckenschmidt. Automating OAEI campaigns (first report). In Asunción Gómez-Pérez, Fabio Ciravegna, Frank van Harmelen, and Jeff Hefflin, editors, *Proc. 1st ISWC international workshop on evaluation of semantic technologies (iWEST), Shanghai (CN)*, page to appear, 2010.
15. Willem Robert van Hage, Antoine Isaac, and Zharko Aleksovski. Sample evaluation of ontology-matching systems. In *Proc. 5th International Workshop on Evaluation of Ontologies and Ontology-based Tools (EON 2007), collocated with ISWC-2007*, pages 41–50, Busan (Korea), 2007.
16. Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In *International Semantic Web Conference*, pages 650–665, 2009.

Gyeongsan, Milano, Mannheim, Trento, Grenoble, Prague, October 2010