Results of the Ontology Alignment Evaluation Initiative 2010*

Jérôme Euzenat[1], Alfio Ferrara[2], Christian Meilicke[3], Andriy Nikolov[4], Juan Pane[5], François Scharffe[1], Pavel Shvaiko[6], Heiner Stuckenschmidt[3], Ondřej Šváb-Zamazal[7], Vojtěch Svátek[7], and Cássia Trojahn[1]

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

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 which provides 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. Our ambition is that from such evaluations, tool developers can improve their systems.

* This paper improves on the "First results" initially published in the on-site proceedings of the ISWC workshop on Ontology Matching (OM-2010). The only official results of the campaign, however, are on the OAEI web site.

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

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 campaign occurred in 2005 at the workshop on Integrating Ontologies held in conjunction with the International Conference on Knowledge Capture (K-Cap) [1]. Starting from 2006 through 2009 the OAEI campaigns were held at the Ontology Matching workshops collocated with ISWC [8; 7; 3; 6]. Finally in 2010, the OAEI results were presented again at the Ontology Matching workshop collocated with ISWC, in Shanghai, China[2].

The main novelty 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, a software infrastructure for automatically executing evaluations, and evaluation campaigns for typical semantic web tools, including ontology matching. 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 the last year because there was not enough organizational resources to guarantee a satisfying evaluation.

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 (§2.1). Then, we present the evaluation environment, which has been used by participants to test their systems and launch their evaluation experiments for the campaign (§2.2). Next, we describe the steps of the OAEI campaign (§2.3-2.6) and report on the general execution of the campaign (§2.7).

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 have 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.

[2] http://om2010.ontologymatching.org

[3] http://www.seals-project.eu

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 organizing conferences (the domain being well understandable for every researcher). Additionally, 'interesting correspondences' are also welcome. Results were evaluated automatically against reference alignments and by data-mining and logical reasoning techniques. Sample of correspondences and 'interesting correspondences' were also evaluated manually.

The directories and thesauri track proposed only web directories this years:

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 4000 elementary tests, and the second is composed by a single test involving matching of two large directories (2854 and 6555 nodes each).

Instance matching (§7): The goal of the instance matching 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. Instance matching is organized in two sub-tasks:

Data interlinking (DI) Participants are requested to re-build the links among the available RDF resources. Reference alignments are provided for each resource as RDF alignments.

OWL data track (IIMB & PR): In the OWL data track, data is provided as OWL individuals according to the RDF/XML format, while reference alignments are provided as RDF alignments. IIMB is divided into test cases and reference alignments are automatically generated by introducing controlled modifications in an initial reference ontology instance. Persons-Restaurants (PR) is a small real data test case where participants are requested to run matching tools against two collections of data concerning persons (person1 and person2) and one collection about restaurants (restaurant1).

The Benchmark, Anatomy and Conference datasets 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 instance matching 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.

This year we had to cancel the VLCR (very large crosslingual resources) task since it had not enough participants to be retained. The Single task modality in the Directory track was also canceled due to lack of resources needed to cross check the reference alignments.

Table 1 summarizes the variation in the results expected from the tests under consideration.

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
di	RDF	=	[0 1]	open	EN
iimb	RDF	=	[0 1]	open	EN
vlcr	SKOS	exact-,	[0 1]	blind	DU+EN
	+OWL	closeMatch		expert	

Table 1. Characteristics of the 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 for 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 [19]. 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 participants for debugging their systems, running their own evaluations and manipulating the results immediately in a direct feedback cycle.

In order to start an evaluation, participants had to specify the URL of the matcher service and the name of the matching system to be evaluated. Then they had to select the evaluation task to be used (Anatomy, Benchmark or Conference). The specified web service is validated by the system (two simple ontologies are used to check if the matcher generates alignments in the correct format). In case of problems, 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.

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. Specifically, results can be visualized and manipulated via an OLAP interface (Figure 1).

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.

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

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

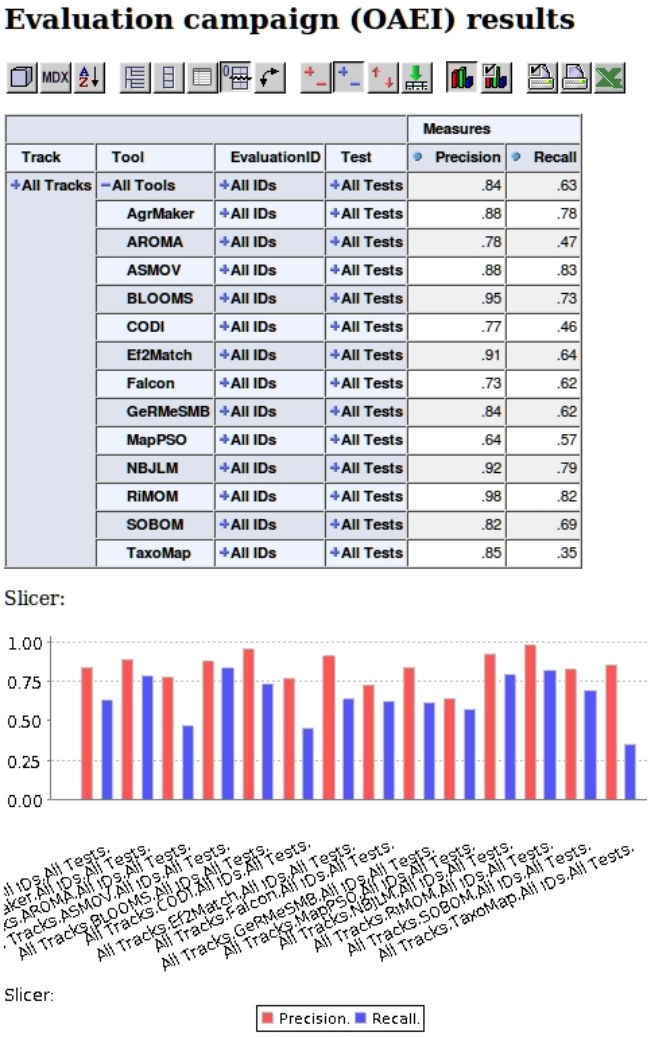


Fig. 1. Using OLAP for results visualization.

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 matched and generate the alignment in the correct format, namely 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 resulting 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 obtained 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 to the fact that the results are automatically computed by the services in the evaluation service and organizers have tools for manipulating and visualizing the results.

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 for compensating 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.

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

The number of covered runs has decreased more than expected: 48 in 2007, 50 in 2008, 53 in 2009, and 37 in 2010. 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 also allow for comparing system on a same execution basis.

The list of participants is summarized in Table 2. Similar to the previous years not all participants provided results for all tests. They usually did those which are easier to run, such as benchmark 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
di												√	√			2
iimb+pr			√		√				√			√	√			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.

Participants may be 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) had 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, 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 is (almost) fully preserved.

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 full results are on the OAEI web site.

As shown in Table 3, two systems achieve top performances: ASMOV and RiMOM, with AgrMaker as a close follower, while SOBOM, GeRMeSMB and Ef2Match, re-

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.
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	0.98	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
Relaxed	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
Weighted	1.00	1.00	0.68	0.57	0.95	0.83	0.94	0.42	0.98	0.61	0.84	0.44	0.98	0.64	0.96	0.46	0.96	0.64	0.86	0.56	0.99	0.83	0.98	0.37	0.87	0.28

Table 3. Results obtained by participants on the benchmark test case (harmonic means). Relaxed precision and recall correspond to the three measures of [4]: symmetric proximity, correction effort and oriented. The same results have been obtained using these three measures. Weighted precision and recall takes into account the confidence associated to correspondence by the matchers.

spectively, had presented intermediary values of precision and recall. In the 2009 campaign, Lily and ASMOV had the best results, with aflood and RiMOM as followers, while GeRoME, AROMA, DSSim and AgrMaker had intermediary performance. The same group of 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. This is because there are no modifications in the labels of classes and properties in these tests and basically all matchers are able to deal with the heterogeneity in labels. For systematic tests (2xx), which allow better to distinguish the strengths of algorithms, ASMOV and RiMOM, respectively, achieve the best results, 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, systems have improved their performance since last year: ASMOV and RiMOM improved their overall performance, AgrMaker and SOBOM significantly improved their recall, while MapPSO and GeRMeSBM improved precision. Only AROMA has significantly decreased in recall. There is no unique set of systems achieving the best results for all cases, which indicates that systems exploiting different features of ontologies perform accordingly to the features of each test case.

The results have also been compared with the relaxed measures proposed in [4], namely symmetric proximity, correction effort and oriented measures ("Relaxed measures" in Table 3). They are different generalisations of precision and recall in order to better discriminate systems that slightly miss the target from those which are grossly wrong. We have used strict versions of these measures (as published in [4] and contrary to previous years). As Table 3 shows, there is no improvement when comparing classical and relaxed precision and recall. This can be explained by the fact that participating algorithms miss the target, by relatively far (the false negative correspondences found by the matchers are not close to the correspondences in the reference alignment) so the gain provided by the relaxed measures has no impact.

We have introduced experimentally confidence-weighted precision and recall in which correspondences are weighted by the confidence matchers put on it. If the confidence is 1., then the correspondence scores exactly like in classical precision and recall. Otherwise, it scores for the amount of confidence. If the correspondence is correct, this will contribute to decrease recall – it will be counted for less than 1. –, if the correspondence is incorrect, this will increase precision – by counting the mistake for less than 1. So this rewards systems able to provide accurate confidence measures (or penalizes less mistakes on correspondences with low confidence). These measures provide precision increase for Falcon, MaPSO and edit distance (which had apparently many incorrect correspondences with low confidence), and recall decrease for Falcon, ASMOV and SOBOM (which had apparently many correct correspondences with low confidence). There are only little variation for other systems. As expected, CODI, which confidence was always 1, shows no variation.

As last year, many algorithms provided 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 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 authorize 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 perform these tests at the highest level. Out of these, ASMOV has slightly better results than the two others. So, this confirms the previous observations on raw results.

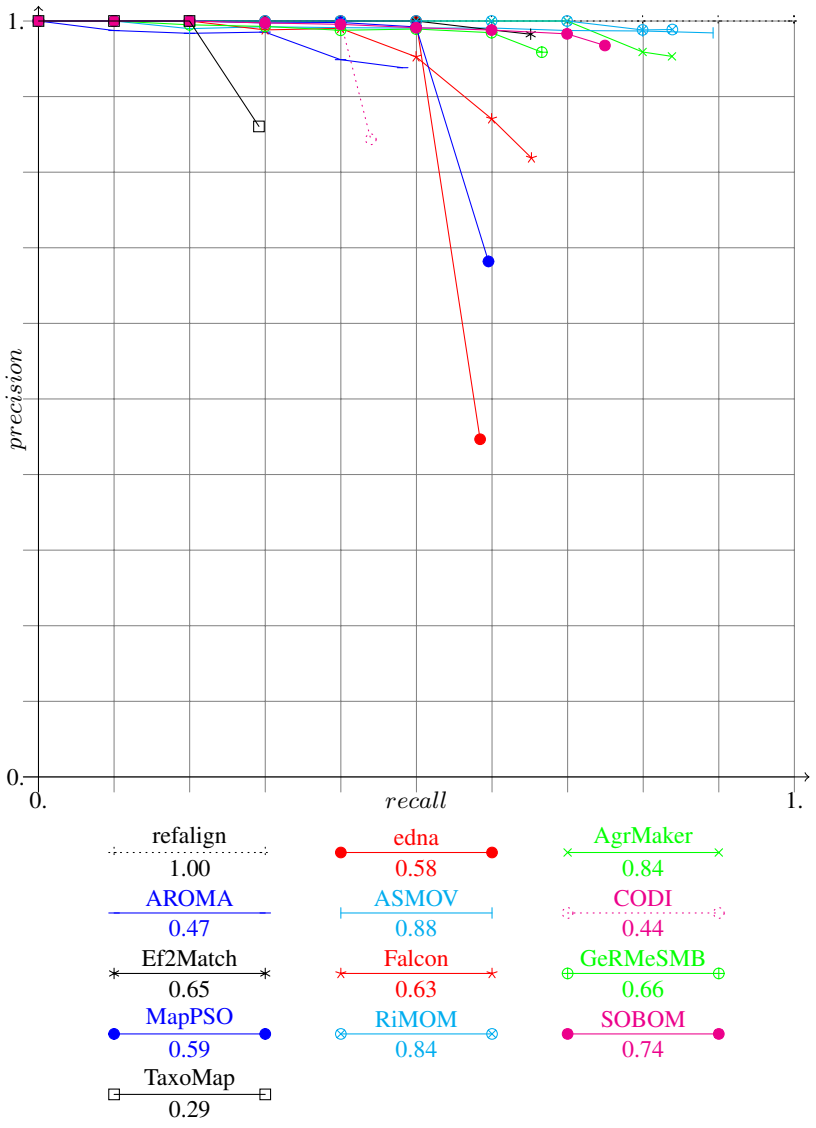


Fig. 2. Precision/recall graphs for benchmarks. The results given by the participants are cut under a threshold necessary for achieving $n\%$ recall and the corresponding precision is computed. Systems for which these graphs are not meaningful (because they did not provide graded confidence values) are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

4 Anatomy

The anatomy track confronts 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. As in previous years, we divided the matching task into four subtasks. Subtask #1 is compulsory 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 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.

We slightly improved the test data set for the 2010 evaluation. We removed some doubtful subsumption axioms and added a number of disjointness statements at the top of the hierarchies to increase the expressivity of the data set. Furthermore, we eliminated some incorrect correspondences.[7]

In previous years, we reported about runtimes 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. 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.

[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 has been roughly stable over four years, we had in 2010 more systems that participated for the first time (5 systems) than in previous years (in average 2 systems; see Table 4 for an overview). Three of the newcomers participate also in other tracks, while NBJLM and BLOOMS participate only in the Anatomy track. Notice also that AgreementMaker (AgrMaker) 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 $\surd$-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% $\rightarrow$ 2008 +5% $\rightarrow$ 2009 +2% $\rightarrow$ 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.

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) generated the best alignment with respect to F-measure. This result is based on a high recall compared to the systems on the following positions. Even the SAMBO system of 2007 could not generate a higher recall with the use of UMLS.

	Task #1			Task #2			Task #3			Recall+	
System	**Prec.**	**F**	**Rec.**	**Prec.**	**F**	**Rec.**	**Prec.**	**F**	**Rec.**	**#1**	**#3**
AgrMaker*	0.903	0.877	0.853	0.962	0.843	0.751	0.771	0.819	0.874	0.630	0.700
Ef2Match	0.955	0.859	0.781	0.968	0.842	0.745	0.954	0.859	0.781	0.440	0.440
NBJLM*	0.920	0.858	0.803	-	-	-	-	-	-	0.569	-
SOBOM	0.949	0.855	0.778	-	-	-	-	-	-	0.433	-
BLOOMS	0.954	0.828	0.731	0.967	0.829	0.725	-	-	-	0.315	-
TaxoMap	0.924	0.824	0.743	0.956	0.801	0.689	0.833	0.802	0.774	0.336	0.414
ASMOV	0.799	0.785	0.772	0.865	0.808	0.757	0.717	0.753	0.792	0.470	0.538
CODI	0.968	0.779	0.651	0.964	0.785	0.662	0.782	0.736	0.695	0.182	0.383
GeRMeSMB	0.884	0.456	0.307	0.883	0.456	0.307	0.080	0.147	0.891	0.249	0.838

Table 5. Results for subtasks #1, #2 and #3 in terms of precision, F-measure, and recall (in addition recall+ for #1 and #3). 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 setting 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 clearly favor precision over recall. 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. This allows to improve a matching system in a direct feedback cycle, however, it might happen that a perfect configuration results in problems for different data sets.

Recall+ and further results. 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, six systems 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 trivial, 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, seven 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 based on a high precision.

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

Table 6. Changes in precision, F-measure and recall 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 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 trend 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 previous years, the input alignment has a characteristics that makes 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.

4.3 Conclusions

Overall, we see a clear improvement comparing 2010 results with the results of 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 reported so far, 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 additionally 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.

[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.

5 Conference

The conference test set introduces matching several moderately expressive ontologies. Within this track, participant results were evaluated using diverse evaluation methods. The evaluation has been supported by the SEALS evaluation service this year.

5.1 Test data

The data set of this track has been extended with one ontology being in OWL 2.[9] For a data set description we refer the reader to the OAEI 2009 results paper [6].

5.2 Results

We had eight participants: AgreementMaker (AgrMaker), AROMA, ASMOV, CODI, Ef2Match, Falcon, GeRMeSMB and SOBOM. All participants delivered all 120 alignments. CODI delivered 'certain' correspondences, the other matchers delivered correspondences with graded confidence values between 0 and 1.

Evaluation based on the reference alignments. We evaluated the results of participants against reference alignments. They include all pairwise combinations between 7 different ontologies, i.e. 21 alignments.

matcher	confidence threshold	Prec.	FMeas.	Rec.
AgrMaker	0.66	.53	.58	.62
AROMA	0.49	.36	.42	.49
ASMOV	0.22	.57	.60	.63
CODI	*	.86	.62	.48
Ef2Match	0.84	.61	.60	.58
Falcon	0.87	.74	.59	.49
GeRMeSMB	0.87	.37	.43	.51
SOBOM	0.35	.56	.56	.56

Table 7. 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 7). Precision, recall, and F-measure are given for this optimal confidence threshold. The dependency of F-measure on the 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 CODI which did not provide graded confidence values. Other matchers are very close to this score (e.g. ASMOV with F-Measure 0.60, Ef2Match with F-Measure 0.60, Falcon with F-Measure 0.59). However, we should take into account that this evaluation has been made over a subset of all alignments (one fifth).

[9] Ontologies have been developed within the OntoFarm project `http://nb.vse.cz/~svatek/ontofarm.html`

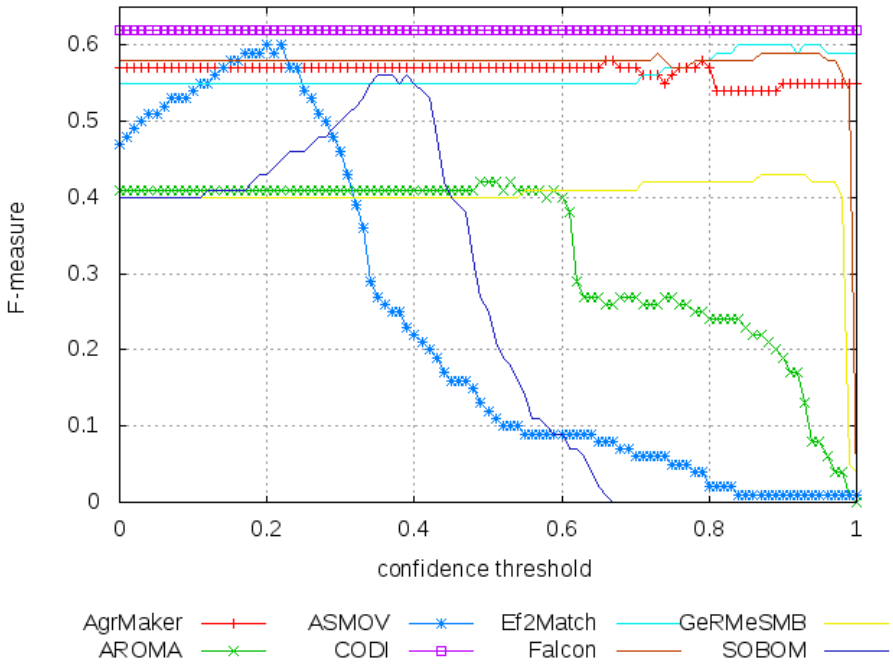


Fig. 3. F-measures depending on confidence.

Comparison with previous years. Three matchers also participated in the last two 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 with .58 in 2010 regarding highest average F-measure. Finally, AROMA participated with the same highest average F-measure in both 2009 and 2010.

Evaluation based on posterior manual labeling. This year we took the most secure correct correspondences, i.e., with the highest confidence, as a population for each matcher. Per matcher, we evaluated 100 correspondences 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 took the 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.

Table 8 presents approximated precisions for each matcher over its population of best correspondences. N is the population of all best correspondences for one matcher. n is a number, ideally 100, of randomly chosen correspondences, among the 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 [20].

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	.68	.86	.98	.79	.96	.30	.82
	± 9.4	± 2.7		± 9.3	± 9.6	± 4.6	± 3.0	± 2.2

Table 8. Approximated precision for 100 best correspondences for each matcher.

From Table 8 we can conclude that CODI, Falcon and AgreementMaker have the best precision (higher than 90%) over their 100 more confident correspondences.

Evaluation based on data mining supported with matching patterns. Results of OAEI participants, i.e. correspondences, contain several attributes characterizing these correspondences from different aspects. Additionally, there is also information in which matching patterns the given correspondence participates; for details about this see [15; 22].

In total there are nine matching patterns ($MP1$ - $MP9$). $MP4$, $MP5$, and $MP6$ are inspired by correspondence patterns from [16]. In principle, it is not possible to say which matching pattern is desirable or not desirable. This must be decided on the basis of an application context or possible alternatives. However, we could roughly say that while $MP2$ and $MP5$ seems to be desirable, $MP7$, $MP8$, and $MP9$ indicate incorrect correspondences related to inconsistency (see section below).

	Antecedent				Succedent	Values	
	System	Certainty factor	Resource1	Resource2	Result	Supp	AvgDff
t1	AgrMaker	$< 1.0; 1.0 >$	*	*	+	0.024	0.95
t2	ASMOV	$< 0.4; 0.8)$	*	*	+	0.01	0.6
t3	SOBOM	< 0.4	*	*	+	0.024	0.37
t4	ASMOV	< 0.4	*	w	-	0.013	1.25
t5	SOBOM	*	*	w	-	0.014	1.21

Table 9. Hypotheses for tasks 1 and 2.

	Antecedent		Succedent	Values
	System	ResultMP	Supp	AvgDff
m1	ASMOV	MP2	0.013	0.79
m2	AgrMaker	MP2	0.015	0.1
m3	AROMA	MP4	0.016	0.16

Table 10. Association hypotheses related to matching patterns.

For data mining we employed the *4ft-Miner* procedure of the *LISp-Miner* data mining system[10] for mining of *association rules*. We found several interesting association

[10] `http://lispminer.vse.cz/`

hypotheses: $t1$ to $t5$ are related to confidence value or underlying resources of ontologies (see Table 9) and $m1$ to $m3$ are related to matching patterns (see Table 10). In total, there were 16522 correspondences in the data matrix. For instance we can interpret *hypothesis t1* as follows *correspondences that are produced by AgrMaker and have the highest confidence value (i.e. 1.0) are by 95% (i.e. almost twice) more often correct than correspondences produced by all systems with all confidence values (on average).*

In conclusion, regarding first three hypotheses we could say that AgrMaker is quite sure about correspondences with the highest value than other matchers. On the other side, ASMOV is surprisingly correct about correspondences with lower confidence values than other matchers. These hypotheses confirmed findings from the previous year since these two matchers also participated in the OAEI-2009. SOBOM is more correct for correspondences with the lowest confidence values. Hypotheses t4 and t5 point out that ASMOV and SOBOM work worse with ontologies based on the web. Regarding hypotheses containing matching patterns, ASMOV and AgrMaker found MP2, while AROMA found MP4 more often than other systems. In comparison with the previous year, there are no interesting hypotheses with matching patterns related to inconsistency. This can be explained by the fact that many occurrences of matching patterns related with inconsistency is generally lower than in the previous year of OAEI. These findings roughly correspond with the results of the evaluation based on the alignment coherence in the next section.

Evaluation based on alignment coherence. In the following we apply the Maximum Cardinality measure as proposed in [14] to measure the degree of alignment incoherence. Results are depicted in Table 11 which shows the average for all testcases of the conference track except the testcases where the ontologies confious and linklings are involved. These ontologies resulted in some combinations of ontologies and alignments in reasoning problems. Note that we did not use the original alignments, but the alignments with optimal threshold. However, the average size of the resulting alignment still varies to a large degree.

Matcher	AgrMaker	AROMA	ASMOV	CODI	Ef2Match	Falcon	GeRMeSMB	SOBOM
Max-Card %	>14.8%	>17.5%	5.6%	0.1%	7.2%	>4.8%	>12.6%	>10.7
N	17.1	16.4	18.2	6.9	12.8	8.9	18.2	11.7

Table 11. Degree of incoherence and size of alignment in average for the optimal a posteriori threshold. The prefix $>$ is added whenever the search algorithm stopped in one of the testcase due to a timeout of 1000 seconds prior to finding the solution.

Compared to the other participants CODI generates the lowest degree of incoherence. This result is partially caused by the small size of alignments that make the occurrence of an incoherence less probable. Taking this into account, the ASMOV system achieves a remarkable result. Even though the alignments of ASMOV comprise the highest number of correspondences, the degree of incoherence 5.6% is relatively small due to the verification component built into the system [12]. Overall it is a suprising result that still only few matching systems take alignment incoherence into account.

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". This task focusses on evaluating the performances of existing matching tools in a real world taxonomy integration scenario.

6.1 Test set

As in previous years, the data set exploited in the directory matching task was constructed from the 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`.

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 are:

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

6.2 Results

In OAEI-2010, 3 out of 15 matching systems participated in 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. The systems that submitted their results to the Single task modality of the Directory track were ASMOV, GeRoMe-SMB, MapPSO and TaxoMap, though the task was canceled due to lack of resources needed to cross check the reference alignments.

Precision, F-measure and recall 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 4.0.

We can observe 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

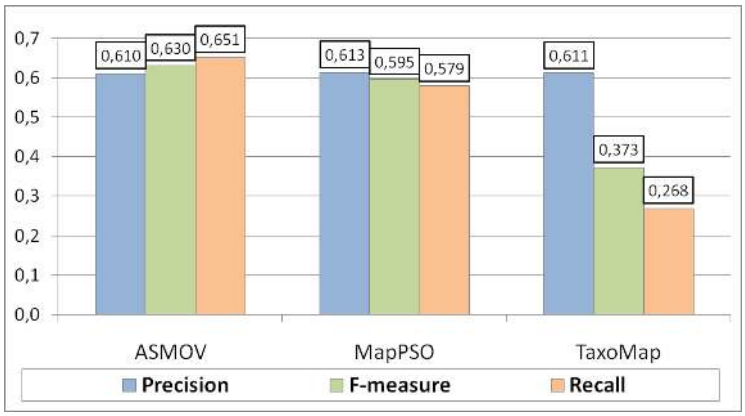


Fig. 4. Matching quality results.

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.

In total, 24 matching systems have participated during the 6 years (2005 - 2010) of the OAEI campaign in the directory track where 40 individual 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).

As can be seen in Figure 5, this year there is an 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 2010 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 does not contain values of precision and F-measure for 2005.

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

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 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) by OLA_2.

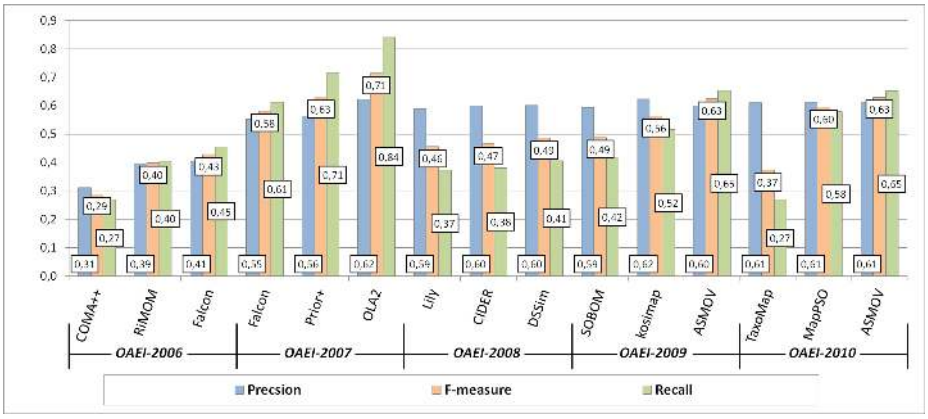
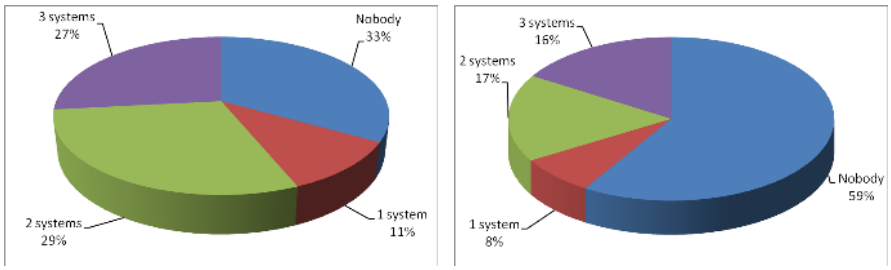


Fig. 6. Comparison of matching quality results in 2006 - 2010.

Partitions of positive and negative correspondences, according to the system results, are presented in Figure 7 **a)** and Figure 7 **b)**, respectively. Figure 7 **a)** 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 7 **b)** 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 7 **b)** 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. The last two observations also suggest that the discrimination ability of the dataset remains as high as in previous years.



a) **b)**

Fig. 7. Partition of the system results: **a)** on positive correspondences, **b)** on negative correspondences.

Figure 7 **a)** 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 of 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 in the last years and the other systems do not seem to cope with the results of 2007.

Figure 7 **b)** 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 being "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 of 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 also 98% of the negative correspondences.

6.3 Comments

This year the average performance of the participants (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 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 7 **a)** and Figure 7 **b)**), 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 gained more importance in the recent years. Unlike the other tracks, the instance matching tests specifically focus on an 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 included two tasks. The first task (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) uses a set of artificially generated and real test cases respectively. These 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 12.

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

Table 12. Participants in the instance matching track.

7.1 Data interlinking task (DI)

Data interlinking is known under many names according to various research communities: equivalence mining, record linkage, object consolidation and coreference 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 becomes more critical. It is particularly important to maximize the automation of the interlinking process in order to be able to follow this expansion.

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 (see [13] 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.

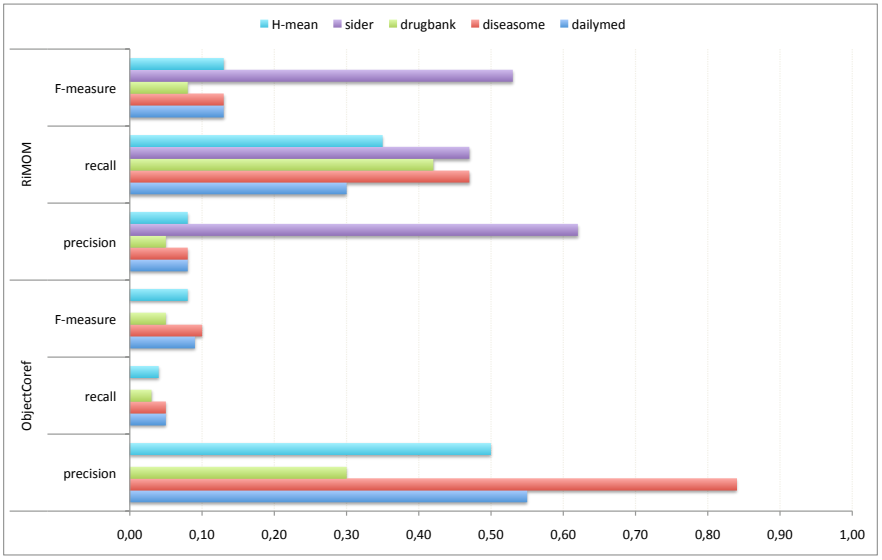


Fig. 8. Results of the DI task.

diseasome contains information about 4300 disorders and genes.

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

These datasets were semi-automatically interlinked using Silk [21] and ODD Linker [11] providing the reference alignments for this task and participants were asked to retrieve these links using an automatic method.

Only two systems participated in the data interlinking task, probably due to the difficulties of matching large collections of data: ObjectCoref and RiMOM. The results of these systems are shown in Figure 8.

The results are very different for two systems, with ObjectCoref being better in precision and RiMOM being better in recall. A difficult task with real interlinked data is to understand if the results are due to a weakness of the matching system or because links can be not very reliable. In any case, what we can conclude from this experiment 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 task (IIMB & PR)

The OWL data task is focused on two main goals:

1. to provide an evaluation dataset for various kinds of data transformations, including value transformations, structural transformations 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 referring to the same objects are provided in different data sources; on the other side, we generate different datasets with a variable level of data quality and complexity. IIMB provides transformation techniques supporting modifications of data property values, modifications of number and type of properties used for the individual description, and modifications of the individuals classification. The first kind of transformations is called *data value transformation* and it aims at simulating the fact that data expressing 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 1416 individuals. In Figures 9 and 10, we report the results over the large version that are quite similar to the small one.

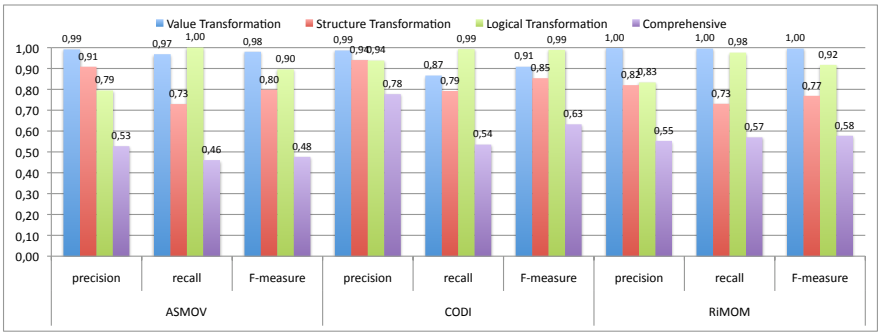


Fig. 9. Results of the IIMB subtrack.

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 worse 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 appear to be comparable in terms of quality of results.

PR. The Person-Restaurants benchmark is composed of three data subsets. 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 with 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. In all the datasets the number of records is quite limited (about 500/600 entries). Among these, 112 record pairs refer to the same entity, but usually show differences. Results of the evaluation are shown in Figure 11.

The PR task of the instance matching track was quite successful in terms of participation, in that all the five systems sent their results for this task[14]. This is because the PR datasets contain a small number of instances to be matched, resulting in a matching task

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

[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

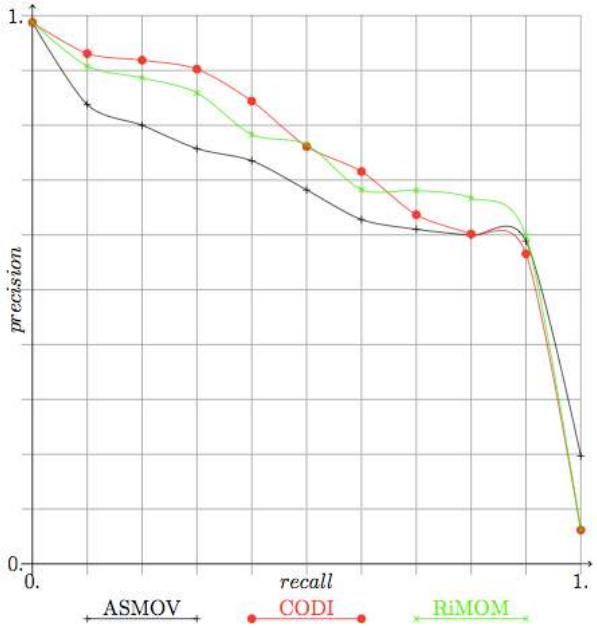


Fig. 10. Precision/recall of tools participating in the IIMB subtrack.

that is affordable in terms of time required for comparisons. The results are good for all systems with the best performances obtained by RiMOM followed by ObjectCoref and LN2R. ASMOV and CODI instead have lower F-measure values in the case of the Person 2 dataset. This is mainly due to low performances in terms of recall. These low recall values 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 service for evaluating ontology matching, the SEALS evaluation service, which has 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 the 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. However, we still have as many participants as last year, so this is a good sign.

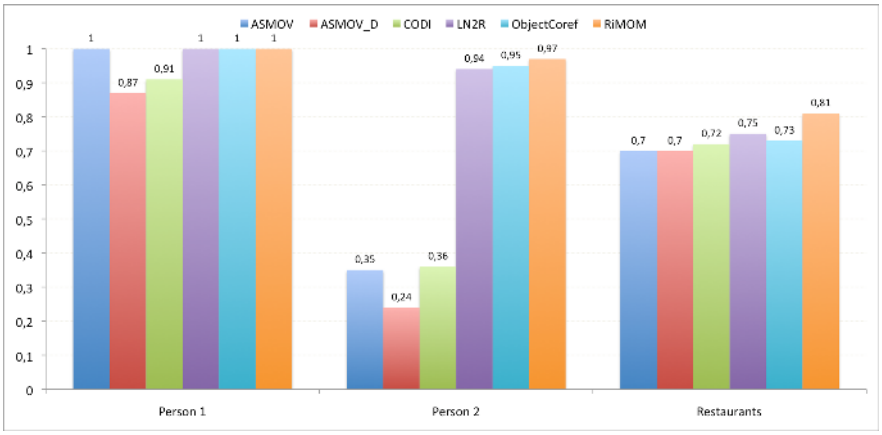


Fig. 11. Results of tools participating in the PR subtrack in terms of F–measure.

B) Although some tools were registered in the SEALS portal, these have not used the SEALS evaluation service either for testing their tools or for registering their final results. We contacted these developers, who answered that they did not have enough time for preparing their tools. So, the effort required for implementing the web service interface and fixing networks issues, has indeed been an obstacle for some participants. However, this effort is low with respect to that required for developing a serious matcher.

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

D) 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. To address the first issue we have identified in [17] the challenges in the field to direct research into the critical path.

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

F*) 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.

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 service. 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 be more challenging for participants. We also plan to generalize the use of the platform to other data sets. Finally, we would like to have again a data set 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 the specialization of systems. While, it is not the result of the short time allowed to the campaign, since the SEALS evaluation service had more runs 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 task 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 for her thorough support on improving the quality of the data set.

We are grateful to Dominique Ritze for participating in the extension of the reference alignments for the conference track.

We thank 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: Yannis Kalfoglou (Ricoh laboratories, UK), Miklos Nagy (The Open University (UK), Natasha Noy (Stanford University, USA), Yuzhong Qu (Southeast University, CN), York Sure (Leibniz Gemeinschaft, DE), Jie Tang (Tsinghua University, CN), George Vouros (University of the Aegean, GR).

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 CSF grant P202/10/1825 (PatOMat project).

References

1. Benhamin Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Integrating Ontologies'05, Proc. of the K-Cap Workshop on Integrating Ontologies*, Banff (Canada), 2005.
2. Oliver Bodenreide, Terry Hayamizu, Martin Ringwald, Sherri De Coronado, and Songmao Zhang. Of mice and men: Aligning mouse and human anatomies. In *Proc. of the 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. of the 3rd International Workshop on Ontology Matching (OM-2008), collocated with ISWC-2008*, pages 73–120, Karlsruhe (Germany), 2008.
4. Marc Ehrig and Jérôme Euzenat. Relaxed precision and recall for ontology matching. In *Proc. of the K-Cap Workshop on Integrating Ontologies*, pages 25–32, Banff (Canada), 2005.
5. Jérôme Euzenat. An API for ontology alignment. In *Proc. of the 3rd International Semantic Web Conference (ISWC-2004)*, pages 698–712, Hiroshima (Japan), 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 *Proc. of the 4th Workshop on Ontology Matching (OM-2009), collocated with ISWC-2000*, pages 73–126, Chantilly (USA), 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. of the 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. of the 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. Yves R. Jean-Mary, E. Patrick Shironoshitaa, and Mansur R. Kabuka. Ontology matching with semantic verification. *Journal of Web Semantics*, 7(3):235–251, 2009.
13. Anja Jentzsch, Jun Zhao, Oktie Hassanzadeh, Kei-Hoi Cheung, Matthias Samwald, and Bo Andersson. Linking open drug data. In *Proc. of Linking Open Data Triplification Challenge at the I-Semantics*, 2009.
14. Christian Meilicke and Heiner Stuckenschmidt. Incoherence as a basis for measuring the quality of ontology mappings. In *Proc. of the 3rd Workshop on Ontology Matching (OM-2008), collocated ISWC-2008*, pages 1–12, Karlsruhe (Germany), 2008.
15. Vojtěch Svátek Ondřej Šváb-Zamazal O. Empirical knowledge discovery over ontology matching results. In *Proc. of the 1st International Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (IRMLeS-2009), collocated with ESWC-2009*, pages 1–15, Heraklion (Greece), 2009.
16. François Scharffe. *Correspondence Patterns Representation*. PhD thesis, University of Innsbruck, 2009.
17. Pavel Shvaiko and Jérôme Euzenat. Ten challenges for ontology matching. In *Proc. of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE-2008)*, pages 1164–1182, Monterrey (Mexico), 2008.
18. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proc. of the Workshop on Evaluation of Ontology-based Tools (EON-2004), collocated with ISWC-2004*, Hiroshima (Japan), 2004.
19. Cássia Trojahn dos Santos, Christian Meilicke, Jérôme Euzenat, and Heiner Stuckenschmidt. Automating OAEI campaigns (first report). In *Proc. of the 1st International Workshop on Evaluation of Semantic Technologies (iWEST-2010), collocated with ISWC-2010*, Shanghai (China), 2010.
20. Willem Robert van Hage, Antoine Isaac, and Zharko Aleksovski. Sample evaluation of ontology-matching systems. In *Proc. of the 5th International Workshop on Evaluation of Ontologies and Ontology-based Tools (EON-2007), collocated with ISWC-2007*, pages 41–50, Busan (Korea), 2007.
21. Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In *Proc. of the 8th International Semantic Web Conference (ISWC-2009)*, pages 650–665, Chantilly (USA), 2009.
22. Ondřej Šváb, Vojtěch Svátek, and Heiner Stuckenschmidt. A study in empirical and 'casuistic' analysis of ontology mapping results. In *Proc. of the 4th European Semantic Web Conference (ESWC-2007)*, pages 655–669, Innsbruck (Austria), 2007.