First results of the Ontology Alignment Evaluation Initiative 2011*

Jérôme Euzenat[1], Alfio Ferrara[2], Willem Robert van Hage[3], Laura Hollink[4], Christian Meilicke[5], Andriy Nikolov[6], François Scharffe[7], Pavel Shvaiko[8], Heiner Stuckenschmidt[5], Ondřej Šváb-Zamazal[9], and Cássia Trojahn[1]

[1] INRIA & LIG, Montbonnot, France
{jerome.euzenat,cassia.trojahn}@inria.fr
[2] Universita degli studi di Milano, Italy
ferrara@dico.unimi.it
[3] Vrije Universiteit Amsterdam, The Netherlands
W.R.van.Hage@vu.nl
[4] Delft University of Technology, The Netherlands
l.hollink@tudelft.nl
[5] University of Mannheim, Mannheim, Germany
{christian,heiner}@informatik.uni-mannheim.de
[6] The Open University, Milton Keynes, UK
A.Nikolov@open.ac.uk
[7] LIRMM, Montpellier, FR
francois.scharffe@lirmm.fr
[8] TasLab, Informatica Trentina, Trento, Italy
pavel.shvaiko@infotn.it
[9] University of Economics, Prague, Czech Republic
ondrej.zamazal@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-2011 builds over previous campaigns by having 4 tracks with 6 test cases followed by 18 participants. Since 2010, the 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 2011 campaign.

* 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 2011 workshop on Ontology Matching (OM-2011) and will be the only official results of the 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 [10; 8]. 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.

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) [12]. 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 2010 the OAEI campaigns were held at the Ontology Matching workshops collocated with ISWC [9; 7; 2; 5; 6]. Finally in 2011, the OAEI results are to be presented again at the Ontology Matching workshop collocated with ISWC, in Bonn, Germany[2].

Since last year, we promote an environment for automatically processing evaluations (§2.2), which has been developed within the SEALS project[3]. This project aims at providing a software infrastructure for automatically executing evaluations, and evaluation campaigns for typical semantic web tools, including ontology matching. 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.

This paper serves as an introduction to the 2011 evaluation campaign 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 3-6 discuss in turn the settings and the results of each of the test cases. Section 7 overviews lessons learned from the campaign. Finally, Section 8 outlines future plans and Section 9 concludes the paper.

2 General methodology

We first present the test cases proposed this year to OAEI participants (§2.1). Then, we present the resources used by participants to test their systems and the execution environment used for running the tools (§2.2). Next, we describe the steps of the OAEI campaign (§2.3-2.5) and report on the general execution of the campaign (§2.6).

2.1 Tracks and test cases

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

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

[2] http://om2011.ontologymatching.org

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

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. This year, we used new systematically generated benchmarks, based on other ontologies than the bibliographic one.

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 the conference task 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.

Oriented alignments (benchmark-subs) :

This track focuses on the evaluation of alignments that contain other relations than equivalences. It provides two datasets of real ontologies taken from a) Academia (alterations of ontologies from the OAEI benchmark series), b) Course catalogs (alterations of ontologies concerning courses in the universities of Cornell and Washington). The alterations aim to introduce additional subsumption correspondences between classes that cannot be inferred via reasoning.

Model matching This dataset compares model matching tools from the Model-Driven Engineering (MDE) community on ontologies. The test cases are available in two formats: OWL and Ecore. The model to be matched have been automatically derived from a model-based repository.

Instance matching (§6): 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): 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.

This year we had to cancel the Oriented alignment and Model matching tracks which have had not enough participation.

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

test	formalism	relations	confidence	modalities	language	SEALS
benchmarks	OWL	=	[0 1]	open	EN	√
anatomy	OWL	=	[0 1]	open	EN	√
conference	OWL-DL	=, <=	[0 1]	blind+open	EN	√
di	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 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 platform

In 2010, participants of the Benchmark, Anatomy and Conference tracks have been asked for the first time to use the SEALS evaluation services: they had to wrap their tools as web services and the tools were executed on the machines of the tool developers [13].

In 2011, tool developers had to implement a simple interface and to wrap their tools in a predefined way including all required libraries and resources. A tutorial for tool wrapping has been provided to the participants. This tutorial described how to wrap a tool and how to use a simple client to run a full evaluation locally. After local test have been conducted successfully, the wrapped tool was uploaded for a test to the SEALS portal.

The wrapped tool was uploaded to the SEALS portal[4] and executed on the SEALS platform by the organisers in a semi-automated way. This approach allowed to measure runtime and ensures the reproducibility of the results for the first time in the history of OAEI. As a side effect, this approach ensures also that a tool is executed with the same setting for all of the three tracks. This was already requested by the organizers in the past years. However, this rule was sometimes not taken into account by participants.

2.3 Preparatory phase

Ontologies to be matched and (where applicable) reference alignments have been provided in advance during the period between May 30^{th} and June 27^{th}, 2011. 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 6^{th}. The data sets did not evolve after this period, except for the reference alignment of the anatomy track to which minor changes have been applied to increase its quality.

2.4 Execution phase

During the execution phase, participants used their systems to automatically match the ontologies from the test cases. In most cases, ontologies are described in OWL-DL

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

and serialized in the RDF/XML format [3]. 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.

Participants can self-evaluate their results either by comparing their output with reference alignments or by using the SEALS client which was able to return measurements such as precision and recall.

2.5 Evaluation phase

Participants have been encouraged to provide (preliminary) results or to upload their wrapped tool in the SEALS portal by September 1^{st}. Organisers evaluated the results and gave feedback to the participants. For the SEALS modality, a full-fledged test on the platform has been conducted by the organizers and problems were reported to the tool developer, until finally an executable version of the tool has been uploaded to the SEALS portal.

Participants were asked to send their final results or upload the final version of their tools by September 23^{th}. Participants also provided the papers that are published hereafter. Note that in the past tool developers had been asked to provide a download link to their tool. For the participants of the SEALS tracks each tool is stored and accessible via the SEALS portal.

As soon as first results were available, these results have been published on the web pages of the track organizers. 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. Additionally, we measured for the first time runtimes for all tracks conducted under the SEALS modality.

2.6 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, 15 in 2010 and 18 in 2011. However, participating systems are now constantly changing.

The number of covered runs has increased more than expected: 48 in 2007, 50 in 2008, 53 in 2009, 37 in 2010, and 53 in 2011. This may be due to the increasing specialization of tests: some systems are specifically designed for instance matching or for anatomy.

This year we were able to run most of the matchers in a controlled evaluation environment, in order to test their portability and deployability. This allowed us comparing systems on a same execution basis. This is also a guarantee that the tested system can be executed out of their particular development environment.

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	CSA	CIDER	CODI	LDOA	Lily	LogMap	MaasMtch	MapEVO	MapPSO	MapSSS	OACAS	OMR	Optima	Serimi	YAM++	Zhishi	Total=18
Confidence	√	√	√	√		√	√	√	√	√	√		√	√		√	√	√	
benchmarks	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√		√		16
anatomy	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√		√		16
conference	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√		√		16
di	√															√		√	3
iimb					√														1
Total	4	3	3	3	4	3	3	3	3	3	3	3	3	3	3	1	3	1	53

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 summary of the results track by track is provided in the following sections. Note also that not each of the tools participating in SEALS modality could generate results for the Anatomy track (see Section 4).

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

This year, we departed from the usual bibliographic benchmark that have been used since 2004. We used a new test generator in order to reproduce the structure of benchmark from different seed ontologies [11].

The systematic benchmark test set is built around a seed 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. 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). This is only used for the initial benchmark.

This year we have generated three different benchmarks against which matchers will be evaluated:

benchmark is the benchmark data set that has been used since 2004. It is used for participants to check that they can run the tests. It also allows for comparison with other systems since 2004. The seed ontology concerns bibliographic references and is inspired freely from BibTeX. It contains 33 named classes, 24 object properties, 40 data properties, 56 named individuals and 20 anonymous individuals.

benchmark2 The Ekaw ontology, one of the ontologies from the conference track, has been used as seed ontology for generating the Benchmark2 data set. It contains 74 classes and 33 object properties. The results with this new data set has been provided to participants after the preliminary evaluation. For the preliminary results, we have been based our evaluation on Benchmark2.

benchmark3 Final results will be evaluated against a data set from yet another seed ontology which has not been disclosed to participants.

Having these three data sets will allow to better evaluate the dependency between the results and the seed ontology. The SEALS platform allows for evaluating matchers against these many data sets automatically.

For all datasets, the reference alignments are still limited: they only match named classes and properties and use the "=" relation with confidence of 1. Full description of these tests can be found on the OAEI web site.

3.2 Results

Sixteen systems have participated in the benchmark track of this year's campaign (see Table 2). From the eleven participants last year, only four participated this year (AgrMaker, Aroma, CODI and MapPSO). On the other hand, we count on ten new participants, while two participants (CIDER and Lily) have been participating in previous campaigns. In the following, we present the evaluation results, both in terms of runtime and compliance in relation to reference alignments.

Portability. 18 systems have been registered in the SEALS portal. One has abandoned due to its many requirements on the platform and an other abandoned silently. So, 16 systems bundled their tools into the SEALS format. Furthermore, we have not been able to run the final versions of OMR and OACAS.

Runtime. This year we have been able to measure the performance of matchers in terms of runtime. We used a Linux Ubuntu 10.04 LTS with 2.26 GHz (2 cores) and 3GB RAM. We have evaluated both the runtime using the local version of the SEALS platform and the client used for participants to test their tools. This is a very preliminary setting for mainly testing the deployability of tools into the SEALS platform. For the near future, evaluations will be fully running into the SEALS platform.

Table 3 presents the runtime required for systems to complete the 102 tests in Benchmark2. We also include some simple edit distance algorithm on labels (edna). Unfortunately, we could not run the CODI matcher under our setting. CODI is able to run under the platform on Windows, but has specific requirements not met yet on the Linux version that has been used. Hence, we were not able to compare CODI's runtime with other systems'.

These measure are temporary because, they are the result of only one run. They however give an idea of the results. The final evaluation will be executed over 5 runs. 13 systems out of 16 were able to run in the platform or client.For the most of the systems, we could observe a constant variation (≈ 2mn) on runtime between platform and client, due to the way it is measured within the platform (i.e., invocations to an external timestamp service take, in average, 0.8s per test). The exceptions are CIDER and MapEVO, which behave very differently in both settings. We are trying to find out an explanation for that and, in principle, this is due to the way the matchers apply their learning and optimization algorithms.

System	edna	AgrMaker	Aroma	CSA	CIDER	CODI	LDOA	Lily	LogMap	MaasMtch	MapEVO	MapPSO	MapSSS	OACAS	OMR	Optima	YAM++
Client runtime (mn)	1	3.1	1	4.8	37.9	x	>10h	9.1	2.4	39.9	16.7	5h	4.9	x	x	>10h	9.9
Platform runtime (mn)	2.6	4.7	2.8	6.2	49.1	x	T	11	4.14	41.6	9.6	T	6.7	E	E	T	11.8
Top five F-measure		√		√		√							√	E	E		√

Table 3. Runtime in Benchmark2. 'T' indicates the systems that could not run in the platform due to some internal timeout; 'x' indicates the tool that could not run in none of the settings; and 'E' indicates that the final version of the tool has some internal error.

Most of the 13 systems are able to complete the 102 alignments in less than 10mn, while CIDER and MaasMatch require around 40mn. While Aroma is the faster matcher, Optima and LDOA completed only part of the task in more than 10 hours and then we stopped the processes. Except for CODI, we could observe that better compliance is not necessary proportional to runtime.

Compliance. Table 4 shows the results, by groups of tests, for the systems able to complete the task successfully in less than 10 hours[5]. We display the results of participants as well as those given by edna (simple edit distance algorithm on labels). The full results are on the OAEI web site.

As shown in Table 4, two systems achieve top performance in terms of F-measure: MapSSS and YAM++, with CODI, CSA and AgrMaker as close followers, respectively. Lily and CIDER had presented intermediary values of precision and recall. All systems achieve a high level of precision and relatively low values of recall. Only MapEVO had a significantly lower recall than edna (with LogMap and MaasMtch with slight lower values), while no system had lower precision. At this time, we could not compare these results with the results from previous campaigns because we have used different benchmarks.

Looking for each group of tests, in simple tests (1xx) all systems have similar performance, excluding CODI. 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. Considering that Benchmark2 has one single test in 1xx, the discriminant category is 2xx, with 101 tests. For this category, the top five systems in terms of F-measure (as stated above) are: MapSSS, YAM++, CODI, CSA and AgrMaker, respectively (CIDER and Lily as followers).

The results have also been compared with the relaxed measures proposed in [4], namely symmetric proximity, correction effort and oriented measures (‘Symmetric’, ‘Effort’, ‘P/R-oriented’ in Table 4). 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 4 shows, these measures provide a uniform and limited improvement to most systems. The exception is MapEVO, which has a considerable improvement in precision. This could be explained by the fact this system misses the target, by not that far (the false negative correspondences found by the matcher are close to the correspondences in the reference alignment) so the gain provided by the relaxed measures has a considerable impact. This may also come from the global optimisation of the system which tends to be globally roughly correct as opposed to locally strictly correct as measured by precision and recall.

We have introduced 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

[5] For this preliminary results, we could not provide full results for tools that require more time to complete the task. Their results will be available in the final version of this paper.

system	**refalign**			**edna**			**AgrMaker**			**Aroma**			**CSA**			**CIDER**			**CODI**		
test	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.
1xx	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	1.00	1.00	0.96	1.00	0.96	0.25	0.19	0.15
2xx	1.00	1.00	1.00	0.50	0.51	0.51	0.98	0.71	0.56	0.93	0.68	0.53	0.82	0.72	0.64	0.89	0.70	0.58	0.94	0.74	0.61
H-mean	1.00	1.00	1.00	0.50	0.51	0.52	0.98	0.71	0.56	0.93	0.68	0.53	0.82	0.73	0.65	0.89	0.70	0.58	0.93	0.73	0.60
Symmetric	1.00	1.00	1.00	0.53	0.53	0.54	0.98	0.71	0.56	0.94	0.68	0.54	0.83	0.73	0.66	0.91	0.71	0.59	0.94	0.73	0.61
Effort	1.00	1.00	1.00	0.53	0.53	0.55	0.98	0.71	0.56	0.94	0.68	0.54	0.84	0.73	0.66	0.90	0.71	0.59	0.93	0.72	0.60
P-oriented	1.00	1.00	1.00	0.56	0.56	0.57	0.98	0.71	0.56	0.95	0.68	0.54	0.84	0.74	0.67	0.92	0.72	0.60	0.95	0.74	0.61
R-oriented	1.00	1.00	1.00	0.56	0.56	0.57	0.98	0.71	0.56	0.95	0.68	0.54	0.84	0.74	0.67	0.92	0.72	0.60	0.95	0.74	0.61
Weighted	1.00	1.00	1.00	0.71	0.60	0.52	0.98	0.71	0.56	0.95	0.64	0.49	0.87	0.58	0.44	0.91	0.68	0.54	0.93	0.73	0.60

	Lily			**LogMap**			**MaasMtch**			**MapEVO**			**MapPSO**			**MapSSS**			**YAM**		
test	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.
1xx	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.99	1.00	1.00	0.99	0.96	0.92	1.00	1.00	1.00	1.00	1.00	1.00
2xx	0.93	0.70	0.57	0.99	0.66	0.49	0.99	0.60	0.43	0.54	0.31	0.21	0.63	0.62	0.61	0.96	0.77	0.64	0.97	0.74	0.60
H-mean	0.93	0.70	0.57	0.99	0.67	0.50	0.99	0.61	0.44	0.55	0.32	0.22	0.64	0.62	0.61	0.96	0.77	0.64	0.97	0.74	0.60
Symmetric	0.93	0.71	0.58	0.99	0.67	0.50	0.99	0.61	0.44	0.63	0.33	0.22	0.66	0.64	0.63	0.97	0.77	0.64	0.97	0.74	0.60
Effort	0.93	0.71	0.58	0.99	0.67	0.50	0.99	0.61	0.44	0.63	0.33	0.23	0.66	0.64	0.63	0.97	0.77	0.64	0.98	0.74	0.60
P-oriented	0.94	0.71	0.58	0.99	0.67	0.50	0.99	0.61	0.44	0.64	0.33	0.23	0.68	0.66	0.65	0.97	0.78	0.65	0.98	0.74	0.60
R-oriented	0.94	0.71	0.58	0.99	0.67	0.50	0.99	0.61	0.44	0.64	0.33	0.23	0.68	0.66	0.65	0.97	0.78	0.65	0.98	0.74	0.60
Weighted	0.95	0.56	0.39	0.99	0.55	0.38	0.99	0.61	0.44	0.75	0.34	0.22	0.76	0.58	0.47	0.96	0.77	0.64	0.99	0.14	0.07

Table 4. 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 (precision and recall). Weighted precision and recall takes into account the confidence associated to correspondence by the matchers.

mistakes on correspondences with low confidence). These measures provide precision increasing for most of the systems, specially edna, MapEVO and MapPSO (which had possibly many incorrect correspondences with low confidence). This shows that the simple edit distance computed by edna is good as confidence measure. The weighted precision and recall for edna could be taken as a decent baseline. It also provides recall decrease specially for CSA, Lily, LogMap, MapPSO and YAM++ (which had apparently many correct correspondences with low confidence). The variation for YAM++ is quite impressive: it does not seems to apply any threshold for filtering its results. Some systems, such as AgrMaker, CODI, MaasMatch and MapSSS, generate all correspondences with confidence = 1, so they have no change.

Many algorithms have provided their results with confidence measures. It is thus possible to draw precision/recall graphs in order to compare them. Figure 1 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). As last year, they 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 4). These new graphs illustrate 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 lower confidence.

3.3 Conclusions

Two new participants have outperformed other participants, including some systems that have been already participated in the campaigns. However, we reported preliminary results.

Within a short period for evaluating the final version of tools, we were not able to compare tools using the original benchmark dataset and the new generated one (Benchmark3). Furthermore, we could not run all the tools under the same basis, using the full SEALS platform setting. For the final version of this paper, we plan to work on these tasks. We plan also to provide more comparative results, especially with regards to how new systems behaviour for the original Benchmark dataset and what was the progress made by systems participating in previous campaigns.

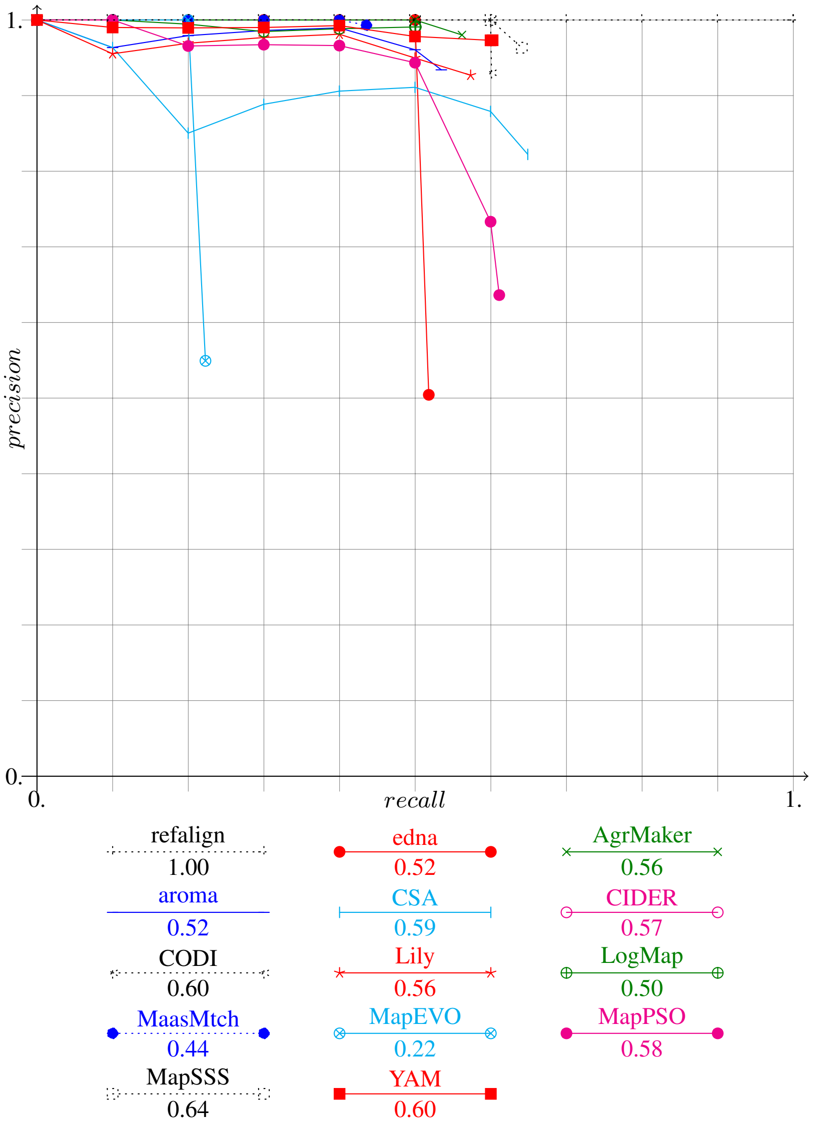

Fig. 1. Precision/recall graphs for benchmarks. The alignments generated by matchers are cut under a threshold necessary for achieving $n\%$ recall and the corresponding precision is computed. The numbers in the legend are the Mean Average Precision (MAP): the average precision for each correct retrieved correspondence. Systems for which these graphs are not meaningful (because they did not provide graded confidence values) are drawn in dashed lines.

4 Anatomy

As in the previous years, the Anatomy track confronts existing matching technology with a specific type of ontologies from the biomedical domain. In this domain, many ontologies have been built covering different aspects of medical research. We focus on (fragments of) two ontologies from this domain, which describe the human anatomy and the anatomy of the mouse. The data set of this track has been used since 2007. For a detailed description we refer the reader to the OAEI 2007 results paper [7].

4.1 Experimental setting

Opposed to the previous years, we distinguish only between two evaluation experiments. Subtask #1 is about applying a matcher with its standard setting to the matching task. In the previous years we have also asked for additional alignments that favor precision over recall and vice versa (subtask #2 and #3). These subtasks are not part of the Anatomy track in 2011 due to the fact that the SEALS platform does not support to run tools with different configurations. Furthermore, we have proposed a fourth subtask, in which a partial reference alignment has to be used as additional input. In the preliminary version of this paper, we do not conduct evaluation experiments related to this specific matching task. We will report about several experiments - with varying input alignments - in the final version of the paper analyzing all tools that support this kind of functionality.

In our experiments we compare precision, recall, F-measure and recall+. We have introduced recall+ to measure the amount of detected non-trivial correspondences. From 2007 to 2009 we reported about runtimes measured by the participants themselves. This survey revealed large differences in runtimes. This year we can compare the runtimes of participants by executing them on our own on the same machine. We used a Windows 2007 machine with 2.4 GHz (2 cores) and 8GB RAM for generating the alignments.

For the 2011 evaluation, we improved again the reference alignment of the data set. We removed doubtful correspondences and included several correct correspondences that had not been included in the past. As a result, we measured for the alignments generated in 2010 a slightly better F-measure (≈+1%) compared to the computation based on the old reference alignment. For that reason we have also included the top-3 systems of 2010 with recomputed precision/recall scores in the results presentation of the following section.

4.2 Results

In the following we analyze the robustness of the submitted systems and their runtimes. Further, we report on the quality of the generated alignment, mainly in terms of precision and recall.

Robustness and Scalability In 2011 there were 16 participants in the SEALS modality, while in 2010 we had only 9 participants for the anatomy track. However, this comparison is misleading. Some of these 16 systems are not really intended to match large biomedical ontologies. For that reason our first interest is related to the question, which systems generate a meaningful result in an acceptable time span. The surprising results are shown in Table 5. First we focused on the question whether the systems finish the matching task in less than 10h. This is the case for a surprisingly low number of systems. The systems that do not finish in time can be separated in those systems that throw an exception related to insufficient memory after some time (marked with 'X', note that we allocated 7GB RAM to the matcher). The other group of systems were still running when we stopped the experiments after 10 hours (marked with 'T').[6]

System	AgrMaker	Aroma	CSA	Cider	CODI	LDOA	Lily	LogMap	MaasMtch	MapEVO	MapPSO	MapSSS	OACAS	OMR	Optima	YAM++
Runtime	634	39	4685	**T**	1890	**T**	563	24	**T**	270	9041	**X**	?	?	**X**	**X**
Meaningful result	√	√	√	-	√	-	√	√	-	-	-	-	-	-	-	-

Table 5. Robustness and performance.

Obviously, matching relatively large ontologies is a problem for six of the 14 systems. The two systems MapPSO and MapEVO (both from the same team) can cope with ontologies that contains more than 1000 concepts, but have problems with finding correct correspondences. MapPSO generates an empty alignment, and MapEVO generates an alignment with precision recall scores less than 5%. This can be related to the way meaningful labels are encoded in the ontologies. The ontologies from the anatomy track differ from the other ontologies in this respect.

For those systems that generate an acceptable result, we observe a high variance in measured runtimes. Clearly ahead is the system LogMap (24 seconds), followed by Aroma (39 seconds). Next are Lily and AgreementMaker (approx. 10 minutes), CODI (30 minutes) and finally CSA (>1h). However, a meaningful interpretation of these runtimes requires to take into account the results for precision and recall.

Results for subtask #1. The results of our experiments are presented in Table 6. Since we have improved the reference alignment, we have also included recomputed precision/recall scores for the top-3 alignments submitted in 2010 (marked by subscript 2010). Keep in mind that in 2010 AgreementMaker (AgrMaker) submitted an alignment that was the best submission to the OAEI anatomy track compared to all previous submissions in terms of F-measure. Note that we also added the base-line *StringEquiv*, which refers to a matcher that compares the normalized labels of two concepts. If these labels are identical, a correspondence is generated. Recall+ is defined as recall, with the difference that the reference alignment is replaced by the set difference of $R \setminus A_{SE}$, where A_{SE} is defined as the alignment generated by *StringEquiv*.

This year we have three systems that generate very good results, namely AgreementMaker, LogMap and CODI. The results of LogMap and CODI are very similar.

[6] The two systems marked with a question mark could not be executed in time. Results will be presented in the final version of this paper. Note also that we will increase the timout to 24 hours (1 day) for the final version of the paper.

Matcher	Precision	F-measure	Recall	Recall+
AgrMaker	0.943	0.917	0.892	0.728
LogMap	0.948	0.894	0.846	0.599
AgrMaker$_{2010}$	0.914	0.890	0.866	0.658
CODI	0.965	0.889	0.825	0.564
NBJLM$_{2010}$	0.931	0.870	0.815	0.592
Ef2Match$_{2010}$	0.965	0.870	0.792	0.455
Lily	0.814	0.772	0.734	0.511
StringEquiv	*0.997*	*0.766*	*0.622*	*0.000*
Aroma	0.742	0.679	0.625	0.323
CSA	0.465	0.576	0.757	0.595

Table 6. Comparison against the reference alignment.

Both systems manage to generate an alignment with F-measure close to the 2010 submission of AgreementMaker. LogMap is slightly ahead. However, in 2011 the alignment generated by AgreementMaker is even better as in the previous year. In particular, AgreementMaker finds more correct correspondences, which can be seen both in recall as well as in recall+ scores. At the same time, AgreementMaker can increase its precision. Also remarkable are the good results of LogMap, given the fact that the system finishes the matching task in less than half a minute. It is thus 25 times faster than AgreementMaker and more than 75 times faster than CODI.

Lily, Aroma, and CSA have less good results than the three top matching systems, however, they have proved to be applicable to larger matching tasks and can generate acceptable results for a pair of ontologies from the biomedical domain. While these systems cannot (or barely) top the String-Equivalence baseline in terms of F-measure, they manage, nevertheless, to generate many correct non-trivial correspondences. A detailed analysis of the results revealed that they miss at the same time many trivial correspondences. This is an uncommon result, which might, for example, be related to some pruning operations performed during the comparison of matchable entities.

4.3 Further Analysis

In the short time span available for generating and analyzing the results, we could not focus on some aspects. In particular, we are also interested in the *coherence* of the generated alignments and the effects of exploiting an *input alignment* (previously known as subtask #4). Both aspects will probably be analyzed in the final version of this paper. Furthermore, we might focus on scalability issues in a set of subsequent experiments (effects of increasing/decreasing number of cores/available RAM).

4.4 Conclusions

Less than half of the systems generate good or at least acceptable results for the matching task of the Anatomy track. With respect to those systems that failed on anatomy, we can assume that the Anatomy track was not in the focus of their developer. This means at the same time that many systems are particularly designed or configured for matching tasks that we find in the benchmark and conference track. Only few of them are robust "all-round" matching systems that are capable of solving different tasks without changing their settings or algorithms.

The positive results of 2011 are the top results of AgreementMaker and the runtime performance of LogMap. AgreementMaker generated a very good result by increasing precision and recall compared to its least years submission, which was clearly the best submission in 2010 already. LogMap clearly outperforms all other systems in terms of runtimes and still generates good results. We refer the reader to the OAEI papers of these two systems for details on the algorithms.

5 Conference

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

5.1 Test data

The collection consists of sixteen ontologies in the domain of organizing conferences. Ontologies have been developed within the OntoFarm project[7].

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 OWL DL axioms of various kinds, which opens a way to use semantic matchers.

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

Participants were to provide all correct correspondences (equivalence and/or subsumption correspondences) and/or 'interesting correspondences' within a collection of ontologies describing the domain of organising conferences.

5.2 Results

This year we newly provide results in terms of F_2measure and $F_{0.5}$measure, comparison with two baselines matchers and precision/recall triangular graph.

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

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	Prec.	F_1Meas.	Rec.	Prec.	F_2Meas.	Rec.	Prec.	$F_{0.5}$Meas.	Rec.
YAM++	.78	**.65**	.56	.78	.59	.56	.8	.73	.53
CODI	.74	.64	.57	.74	.6	.57	.74	.7	.57
LogMap	.84	.63	.5	.84	.54	.5	.85	**.75**	.5
AgrMaker	.65	.62	.59	.58	**.61**	.62	.8	.69	.44
$BaseLine_2$	**.79**	**.59**	**.47**	**.79**	**.51**	**.47**	**.79**	**.7**	**.47**
MassMtch	.83	.56	.42	.83	.47	.42	.83	.69	.42
$BaseLine_1$	**.8**	**.56**	**.43**	**.8**	**.47**	**.43**	**.8**	**.68**	**.43**
CSA	.5	.55	.6	.5	.58	.6	.61	.58	.47
CIDER	.64	.53	.45	.38	.48	.51	.67	.61	.44
MapSSS	.55	.51	.47	.55	.48	.47	.55	.53	.47
Lily	.36	.41	.47	.37	.45	.47	.48	.42	.27
AROMA	.35	.4	.46	.35	.43	.46	.35	.37	.46
Optima	.25	.35	.57	.25	.45	.57	.25	.28	.57
MapPSO	.21	.23	.25	.12	.26	.36	.28	.25	.17
LDOA	.1	.17	.56	.1	.29	.56	.1	.12	.56
MapEVO	.15	.04	.02	.02	.02	.02	.27	.08	.02

Table 7. The highest $F_{[1|2|0.5]}$measure and their corresponding precision and recall for some threshold for each matcher.

For a a better comparison, we established the highest average F-measures for each algorithm. We used F_1measure, which is the harmonic mean of precision and recall. Furthermore, we used F$_2$measure (for $\beta = 2$) which weights recall higher than precision and F$_{0.5}$measure (for $\beta = 0.5$) which weights precision higher than recall. In Table 7 we can see the highest F_1measure, F_2measure and $F_{0.5}$measure and their corresponding precision and recall for some threshold for each matcher.

Matchers are ordered according to their highest F_1measure. Additionally, there are two simple string matchers as baselines. $Baseline_1$ is a string matcher based on string equality applied on local names of entities which were lowercased before. $Baseline_2$ enhanced $baseline_1$ with three string operations: removing of dashes, underscore and ”has” words from all local names. These two baselines divide matchers into four groups. Group 1 consists of best matchers (YAM++, CODI, LogMap and AgreementMaker) having better results than $baseline_2$ in terms of F_1measure. Matchers which perform worse than $baseline_2$ in terms of F_1measure but still better than $baseline_1$ are in Group 2 (MaasMatch). Group 3 (CSA, CIDER and MapSSS) contains matchers which are better than $baseline_1$ at least in terms of F_2measure. Other the matchers (Lily, AROMA, Optima, MapPSO, LDOA and MapEVO) perform worse than $baseline_1$ (Group 4). Optima, MapSSS and CODI did not provide graded confidence values. Performance of matchers regarding F$_1$measure is visualized in Figure 2.

In conclusion, all best matchers (group one) are very close to each other. However, the matcher with the highest average F_1measure (.65) is YAM++, the highest average F_2measure (.61) is AgreementMaker and the highest average $F_{0.5}$measure (.75) is

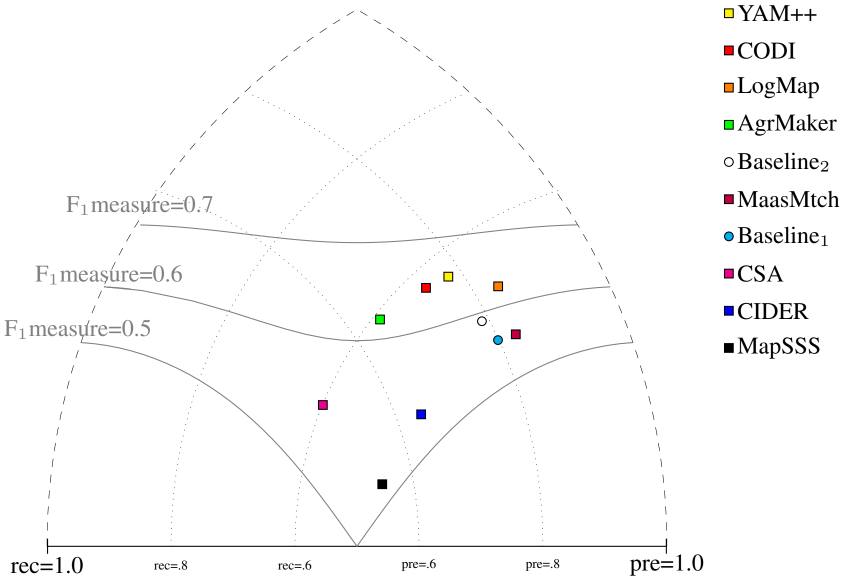

Fig. 2. Precision/recall triangular graph for conference. Matchers of participants from first three groups are represented as squares. Baselines are represented as circles. Horizontal line depicts level of precision/recall while values of F$_1$measure are depicted by areas bordered by corresponding lines F$_1$measure=0.[5|6|7].

LogMap. In any case, we should take into account that this evaluation has been made over a subset of all possible alignments (one fifth).

Comparison with previous years. Three matchers also participated in the previous year. AgreementMaker improved its F_1measure from .58 to .62 by higher precision (from .53 to .65) and lower recall (from .62 to .59), CODI increased its F_1measure from .62 to .64 by higher recall (from .48 to .57) and lower precision (from .86 to .74). AROMA (with its AROMA variant) slightly decreased F_1measure from .42 to .40 by lower precision (from .36 to .35) and recall (from .49 to .46).

6 Instance matching

The instance matching track was included into the OAEI campaigns for the third 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 linked data, 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. 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) 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 8.

Dataset	**AgrMaker**	**SERIMI**	**Zhishi**	**CODI**
DI-nyt-dbpedia-locations	√	√	√	
DI-nyt-dbpedia-organizations	√	√	√	
DI-nyt-dbpedia-people	√	√	√	
DI-nyt-freebase-locations	√	√	√	
DI-nyt-freebase-organizations	√	√	√	
DI-nyt-freebase-people	√	√	√	
DI-nyt-geonames	√	√	√	
IIMB				√

Table 8. Participants in the instance matching track.

6.1 Data interlinking task (DI) – New York Times

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 the task consists of matching the New York Times subject headings to DBpedia, Freebase and Geonames.

The New York Times has developed over the past 150 years an authoritative vocabulary for annotating news items. The vocabulary contains about 30,000 subject headings, or tags. They are progressively published as linked open data and, by July 2010, over 10,000 of these subject headings, in the categories People, Organizations, Locations and Descriptors, have been published[8]. For further information about the New York Times data we refer to their website on linked data Documents.

The New York Times dataset was used in the OAEI campaign for the second time. However, the set of reference links has been updated to reflect the changes made to the external datasets during the year. In particular, several missing links were added, links pointing to non-existing DBPedia instances were removed, and links to instances redirecting to others were updated. Moreover, the Descriptors facet has been removed from the evaluation, since there was not a clear identity criterion for its instances.

Facet	**# Concepts**	**Links to Freebase**	**Links to DBPedia**	**Links to Geonames**
People	4,979	4,979	4,977	0
Organizations	3,044	3,044	1,965	0
Locations	1,920	1,920	1,920	1,920

The SKOS representation of each subject heading facet contains the label of the skos:Concept (skos:label), the facet it belongs to (skos:inScheme), and some specific properties: nyt:associated_article_count for the number of NYT articles the concept is associated with and nyt:topicPage pointing to the topic page (in HTML) gathering different information published on the subject. The concepts have links to DBpedia, Freebase and/or GeoNames. The Location facet also contains geo-coordinates.

DI results An overview of the Precision, Recall and F_1-measure results per dataset of the DI subtrack is shown in table 9. A Precision-Recall graph visualization is shown in figure 3. The results show a variation in both systems and datasets. Zhishi.links produces consistently high quality matches over all datasets, and get's the highest overall scores. Matches to DBpedia locations (DI-nyt-dbpedia-loc.) appear to be difficult as Agreementmaker and SERIMI perform poorly on both precision and recall. This is not the case for Freebase locations (DI-nyt-freebase-loc.) and to a much lesser extent for Geonames (DI-nyt-geonames). We hypthesise that this is due to many locations not being present in dbpedia. Agreementmaker's scores considerably higher on People than on Locations and Organizations, which can be observed in both the DBpedia and the Freebase dataset.

[8] http://data.nytimes.com/

	AgreementMaker			SERIMI			Zhishi.links		
Dataset	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.
DI-nyt-dbpedia-loc.	0,79	0,69	0,61	0,69	0,68	0,67	0,92	0,92	0,91
DI-nyt-dbpedia-org.	0,84	0,74	0,67	0,89	0,88	0,87	0,90	0,91	0,93
DI-nyt-dbpedia-peo.	0,98	0,88	0,80	0,94	0,94	0,94	0,97	0,97	0,97
DI-nyt-freebase-loc.	0,88	0,85	0,81	0,92	0,91	0,90	0,90	0,88	0,86
DI-nyt-freebase-org.	0,87	0,80	0,74	0,92	0,91	0,89	0,89	0,87	0,85
DI-nyt-freebase-peo.	0,97	0,96	0,95	0,93	0,92	0,91	0,93	0,93	0,92
DI-nyt-geonames.	0,90	0,85	0,80	0,79	0,80	0,81	0,94	0,91	0,88
H-mean.	0,92	0,85	0,80	0,89	0,89	0,88	0,93	0,92	0,92

Table 9. Results of the DI subtrack.

6.2 OWL data task (IIMB)

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 the ISLab Instance Matching Benchmark (IIMB). 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 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 transformations 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 transformations, called *data semantic transformation*, simulates the fact that the same real object may be classified in different ways in different data sources.

The 2011 edition of IIMB is created by extracting data from Freebase, an open knowledge base that contains information about 11 million real objects including

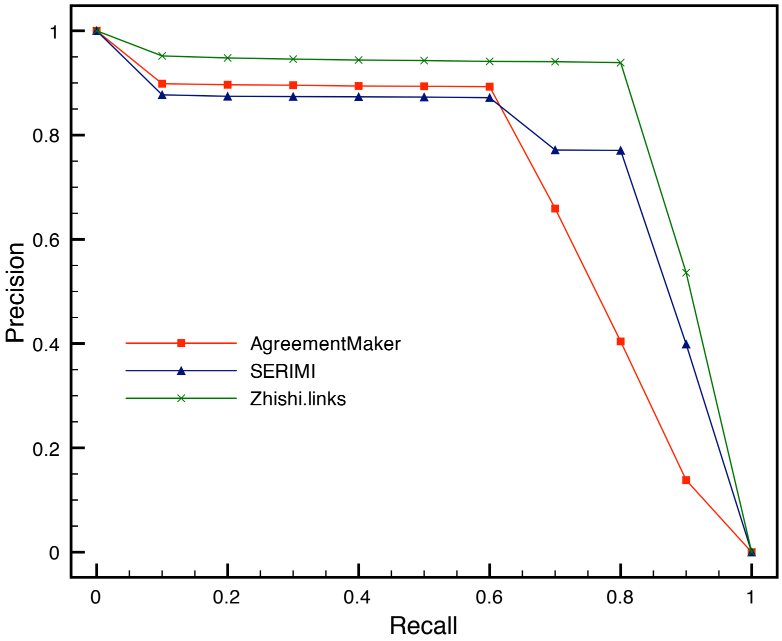

Fig. 3. Precision/recall of tools participating in the DI subtrack.

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[9]. IIMB2011 is a collection of OWL ontologies consisting of 29 concepts, 20 object properties, 12 data properties and thousands of individuals (4000+) divided into 80 test cases.

Test cases from 0 to 20 contain changes in data format (misspelling, errors in text, etcetera); test cases 21 to 40 contain changes in structure (properties missing, RDF triples changed); 41 to 60 contain logical changes (class membership changed, logical errors); finally, test cases 61 to 80 contain a mix of the previous. One system participated in this task. The results of Codi in table 4 show how precision drops moderately and recall drops dramatically as more errors are introduced.

IIMB results An overview of the Precision, Recall and F_1-measure results per set of tests of the IIMB subtrack is shown in table 4. A Precision-Recall graph visualization is shown in figure 5.

7 Lesson learned and suggestions

This year we implemented most of our 2010 future plans by providing a common platform on which evaluation could be performed. There still remain one lesson not really

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

	codi		
test	Prec.	FMeas.	Rec.
001–010	0.94	0.84	0.76
011–020	0.94	0.87	0.81
021–030	0.89	0.79	0.70
031–040	0.83	0.66	0.55
041–050	0.86	0.72	0.62
051–060	0.83	0.72	0.64
061–070	0.89	0.59	0.44
071–080	0.73	0.33	0.21

Fig. 4. Results of the IIMB subtrack.

taken into account that we identify with an asterisk (*) and that we will tackle this year. The main lessons from this year are:

A) This year again as shown that requiring participants to implement a minimal interface was not a strong obstacle to participation. The interface allows for comparing matchers on the same or similar hardware. It also allows for running more tests or reproducing results without running a new campaign.

B) By using the SEALS platform, we have eliminated the network issue that we had last year with web services and we can better testify of the portability of tools.

C) The client available for testing and evaluating wrapped tools was intensively used for participants to test and improve their systems. So, interoperability and the ability to get immediate feedback is appreciated by implementers. Moreover, participants could use the client to generate preliminary results to be included in their papers.

D) Last years we reported that there are not many new systems entering the competition. This year we had many new participants. Only a minority of systems participated in one of the previous years.

E) In spite of claims that such evaluations were needed, we had to declare the Model matching and oriented Alignment tracks unfruitful. This is a pity, but this shows that setting up a data set is not sufficient for getting participants.

F) More interesting, there are only a few matchers involved in the instance matching track. This is especially surprising given the high number of papers submitted and published on this topic nowadays. It seems that people involved in instance matching should cooperate to propose standard formats and evaluation modalities that everyone would use.

G) There is a high variance in runtimes and there seems to be no correlation between runtime and quality of the generated results.

*H) The low number of systems that could generate results for the Anatomy track is an uncommon result. It seems that not many matching systems are capable of matching larger ontologies (>1000 concepts). Even if we had introduced new benchmark generation facilities, we have not used it towards more discriminating benchmarks. This is a topic for immediate future works that we plan to address in the next few months.

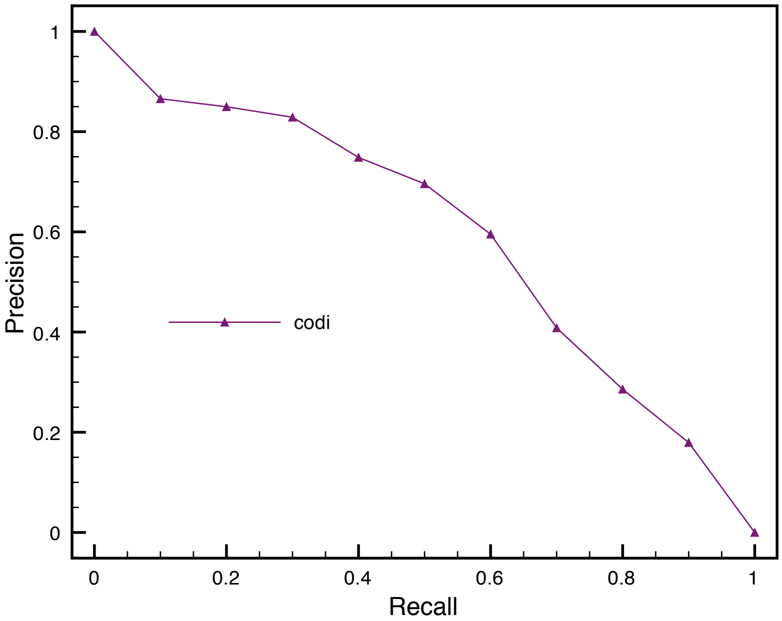

Fig. 5. Precision/recall of CODI tool participating in the IIMB subtrack.

8 Future plans

In 2012, for logistic reasons, we plan to have an intermediate evaluation before OAEI-2012. This evaluation will be concentrated towards exploiting fully the SEALS platform and, in particular:

- performing benchmark scalability tests by reducing randomly a large seed ontology;
- generating discriminating benchmarks by suppressing easy tests;
- adding new tasks, such as multilingual conferences, on the SEALS platform.

We plan to run these tests within the next six months with the already registered tools that would like to be evaluated as well as with new tools willing to enter. These partial results will be integrated within the results of OAEI-2012.

9 Conclusions

Confirming the trend of previous years, the number of systems, and tracks they enter in, seems to stabilize. The trend of number of tracks entered by participants went down again: 2.9 against 2.6 in 2010, 3.25 in 2009, 3.84 in 2008 and 2.94 in 2007. This figure of around 3 out of 7 may be the result of the specialization of systems. This number is dominated by the use of the SEALS platform: each tool entering there can be evaluated on three tasks.

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 matching tool executable in time and they provided insightful papers presenting their experience. The best way to learn about the results remains to read the following papers.

We are grateful to 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. 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 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 has been 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. 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.

3. Jérôme David, Jérôme Euzenat, François Scharffe, and Cássia Trojahn dos Santos. The alignment api 4.0. *Semantic web journal*, 2(1):3–10, 2011.
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, 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.
6. Jérôme Euzenat, Alfio Ferrara, Christian Meilicke, Andriy Nikolov, Juan Pane, François Scharffe, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, and Cássia Trojahn dos Santos. Results of the ontology alignment evaluation initiative 2010. In Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Ming Mao, and Isabel Cruz, editors, *Proc. 5th ISWC workshop on ontology matching (OM), Shanghai (CN)*, pages 85–117, 2010.
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, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, and Cássia Trojahn dos Santos. Ontology alignment evaluation initiative: six years of experience. *Journal on Data Semantics*, XV:158–192, 2011.
9. 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.
10. Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer, Heidelberg (DE), 2007.
11. Maria Rosoiu, Cássia Trojahn dos Santos, and Jérôme Euzenat. Ontology matching benchmarks: generation and evaluation. In Pavel Shvaiko, Isabel Cruz, Jérôme Euzenat, Tom Heath, Ming Mao, and Christoph Quix, editors, *Proc. 6th ISWC workshop on ontology matching (OM), Bonn (DE)*, 2011.
12. 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.
13. 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.

Grenoble, Milano, Amsterdam, Delft, Mannheim, Milton-Keynes, Montpellier, Trento,
Prague, September 2011