HAL open science

Results of the first evaluation of matching tools

Cassia Trojahn dos Santos, Christian Meilicke, Jérôme Euzenat, Ondrej Sváb-Zamazal

To cite this version:

Cassia Trojahn dos Santos, Christian Meilicke, Jérôme Euzenat, Ondrej Sváb-Zamazal. Results of the first evaluation of matching tools. [Contract] 2010, pp.36. hal-00793288

HAL Id: hal-00793288

https://inria.hal.science/hal-00793288v1

Submitted on 22 Feb 2013

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

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

SEALS

Semantic Evaluation at Large Scale

FP7 – 238975

D12.3 Results of the first evaluation of matching tools

Coordinator: Cássia Trojahn
With contributions from: Christian Meilicke, Jérôme Euzenat, Ondřej Šváb-Zamazal
Quality Controller: Heiner Stuckenschmidt
Quality Assurance Coordinator: Raúl García Castro

Document Identifier:	SEALS/2010/D12.3/V1.0
Class Deliverable:	SEALS EU-IST-2009-238975
Version:	version 1.0
Date:	November 25, 2010
State:	final
Distribution:	public

Executive Summary

The first SEALS evaluation campaign has been conducted in coordination with the OAEI 2010 campaign. The Ontology Alignment Evaluation Initiative (OAEI) is a coordinated international initiative that organizes annual evaluation campaigns of ontology matching systems. The campaigns have been carried out since 2004. Usually, matcher developers run their tools in their own machines and submit their results (alignments) to the organizers, that are responsible for evaluating them. It is a time consuming task that requires a lot of communication between developers and organizers. Furthermore, it involves checking if alignments are in the correct format (and eventually correcting them) and running scripts for measuring the quality of the generated alignments.

The first effort in automating the OAEI campaigns has been done in cooperation with the SEALS project. The aim is to integrate progressively the SEALS infrastructure within the OAEI campaigns. For the OAEI 2010 campaign, some of the OAEI tracks have been included in a new modality, the SEALS modality, while for other tracks the conventional process has been conducted.

In this deliverable, we report the results of the first OAEI/SEALS integrated campaign, focusing on the SEALS modality. First, we describe the subset of the OAEI tracks (§2.1) that has been included in the new modality, namely Anatomy, Benchmark and Conference. The datasets have been selected because they are well known to the organizers and have been used in many evaluations. Furthermore, they come with a high quality reference alignment, what allows for evaluating the compliance of generated alignments (§2.2). We have focused on standard metrics, such as precision and recall. For the Conference test set, only a partial reference alignment is available and for this reason we have applied additionally manual labeling and measured the degree of alignment incoherence. We comment on how the evaluation process has been conducted for the SEALS tracks (§2.3). Finally, we list the campaign participants (§2.4).

From the participant's point of view, the main innovation is the use of a web-based user interface for executing evaluations. We briefly illustrate how to run a complete evaluation cycle using this user interface (§3). The preliminary evaluation results are then presented (§4) for each SEALS track. There is no unique set of systems ahead for all three tracks, what clearly demonstrates that systems exploit different features of ontologies and perform accordingly to the features of each test cases. Finally, we comment on the main learned lessons (§5), highlighting the major benefits and limitations of the integration of the evaluation service into the OAEI 2010 campaign.

Document Information

IST Project Number	FP7 – 238975	**Acronym**	SEALS
Full Title	Semantic Evaluation at Large Scale		
Project URL	`http://www.seals-project.eu/`		
Document URL			
EU Project Officer	Carmela Asero		

Deliverable	**Number**	12.3	**Title**	Results of the first evaluation of matching tools
Work Package	**Number**	12	**Title**	Matching Tools

Date of Delivery	**Contractual**	M18	**Actual**	20-10-10
Status	version 1.0		final ☒	
Nature	prototype ☐ report ☒ dissemination ☐			
Dissemination level	public ☒ consortium ☐			

Authors (Partner)	Christian Meilicke (University Mannheim), Jérôme Euzenat, Ondřej Šváb-Zamazal (University of Economics, Prague, Czech Republic)			
Resp. Author	**Name**	Cássia Trojahn	**E-mail**	cassia.trojahn@inrialpes.fr
	Partner	INRIA	**Phone**	+33 (476) 615 476

Abstract (for dissemination)	This deliverable reports the results of the first SEALS evaluation campaign, which has been carried out in coordination with the OAEI 2010 campaign. A subset of the OAEI tracks has been included in a new modality, the SEALS modality. From the participant's point of view, the main innovation is the use of a web-based interface for launching evaluations. 13 systems, out of 15 for all tracks, have participated in some of the three SEALS tracks. We report the preliminary results of these systems for each SEALS track and discuss the main lesson learned from to the use of the new technology for both participants and organizers of the OAEI.

Keywords	ontology matching, ontology alignment, evaluation, benchmarks

Version Log			
Issue Date	**Rev No.**	**Author**	**Change**
20/10/2010	1	Cassia Trojahn	Set up overall structure
25/10/2010	2	Cassia Trojahn	Added content each section (first version)
26/10/2010	3	Cassia Trojahn	Modified abstract
26/10/2010	4	Cassia Trojahn	Added executive summary
27/10/2010	5	Cassia Trojahn	Added lessons learned (first version)
28/10/2010	6	Cassia Trojahn	Modified evaluation service section
02/11/2010	7	Christian Meilicke	Modified results sections conference/anatomy
03/11/2010	8	Christian Meilicke	Added coherence analysis
03/11/2010	9	Christian Meilicke	Lessons learned section finished
03/11/2010	10	Christian Meilicke	Included updated results table for conference
04/11/2010	11	Christian Meilicke	Final revision of whole document

Project Consortium Information

Participant's name	Partner	Contact
Universidad Politécnica de Madrid		Asunción Gómez-Pérez Email: asun@fi.upm.es
University of Sheffield		Fabio Ciravegna Email: fabio@dcs.shef.ac.uk
Forschungszentrum Informatik		Rudi Studer Email: studer@aifb.uni-karlsruhe.de
University of Innsbruck		Barry Norton Email: barry.norton@sti2.at
Institut National de Recherche en Informatique et en Automatique		Jérôme Euzenat Email: Jerome.Euzenat@inrialpes.fr
University of Mannheim		Heiner Stuckenschmidt Email: heiner@informatik.uni-mannheim.de
University of Zurich		Abraham Bernstein Email: bernstein@ifi.uzh.ch
Open University		John Domingue Email: j.b.domingue@open.ac.uk
Semantic Technology Institute International		Alexander Wahler Email: alexander.wahler@sti2.org
University of Oxford		Ian Horrocks Email: ian.horrocks@comlab.oxford.ac.uk

Table of Contents

List of figures 6

List of tables 7

1 Introduction 8

2 Evaluation Campaign 9
2.1 Test data . . . 9
2.1.1 Benchmark test data . . . 9
2.1.2 Anatomy test data . . . 10
2.1.3 Conference test data . . . 11
2.2 Evaluation criteria and metrics . . . 12
2.3 General methodology . . . 12
2.3.1 Preparatory phase . . . 13
2.3.2 Preliminary tests . . . 13
2.3.3 Execution phase . . . 13
2.3.4 Evaluation phase . . . 14
2.4 Participants . . . 14

3 Running and Registering an Evaluation 16

4 Evaluation Results 22
4.1 Benchmark results . . . 22
4.2 Anatomy . . . 26
4.3 Conference . . . 29

5 Lessons Learned 32

6 Final Remarks 34

References 34

List of Figures

3.1 Specifying a matcher endpoint as evaluation target. 16
3.2 Example of error message displayed to the user. 17
3.3 Listing of available evaluation results for the specific endpoint. 17
3.4 Display results of an evaluation. 18
3.5 View on an alignment. 18
3.6 Using OLAP for results visualization. 19
3.7 Registering the results for the campaign. 19
3.8 List of all registered tools for the campaign. Organizers can see the results for that tool by clicking at its name link. 20
3.9 List of all registered results for all SEALS tracks. 21

4.1 Precision/recall graphs for benchmarks. 25
4.2 F-measures depending on confidence. 30

List of Tables

2.1 Participants and the state of their submissions. 14

4.1 Benchmark results. 23
4.2 Overview on anatomy participants from 2007 to 2010, a $\surd$indicates that the system participated, + indicates that the system achieved an F-measure ≥ 0.8 in subtask #1. 26
4.3 Results for subtasks #1, #2 and #3 in terms of precision, recall (in addition recall+ for #1 and #3) and F-measure. 27
4.4 Changes in precision, recall and F-measure based on comparing $A_1 \cup R_p$ and A_4 against reference alignment R. 28
4.5 Confidence threshold, precision and recall for optimal F-measure for each matcher. 29
4.6 Approximated precision for 100 best correspondences for each matcher. 31
4.7 Degree of incoherence and size of alignment in average for the optimal a-posteriori threshold. 31

1. Introduction

The Ontology Alignment Evaluation Initiative[1] (OAEI) is a coordinated international initiative that organizes the evaluation of ontology matching systems [7]. 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. The ambition is that from such evaluations, tool developers can learn and improve their systems. The OAEI annual campaign provides the evaluation of matching systems on consensus test cases, which are organized by different groups of researchers. OAEI evaluations have been carried out since 2004.

Although OAEI has been the basis for ontology matching evaluation over the last years, additional efforts have to be made in order to catch up with the growth of ontology matching technology, specially in two main directions: large scale evaluation and automation of the evaluation process. To that extent, SEALS aims at providing standardized data sets, evaluation campaigns for typical semantic web tools and, in particular, a software infrastructure for automatically executing evaluations. The SEALS infrastructure will allow developers to run their tools on an execution environment in both the context of an evaluation campaign and on their own for a formative evaluation of their tool versions.

OAEI and SEALS are closely coordinated and the plan is to integrate progressively the SEALS infrastructure within the OAEI campaigns. The 2010 OAEI campaign is the first effort in this direction. A subset of the OAEI tracks have been included in the new modality (SEALS modality). As detailed in [10], participants are invited to extend a web service interface and deploy their matchers as web services, which are accessed in an evaluation experiment. This setting enables participants to debug their systems, run their own evaluations and manipulate the results immediately in a direct feedback cycle.

In this deliverable, we report the results of the first OAEI/SEALS integrated campaign. We focus on the preliminary results for the systems participating in each SEALS track. Furthermore, we comment on the impact of using the new technology in the evaluation process as well as we present the main learned lessons, highlighting the major benefits and limitations of the integration of the evaluation service into the OAEI 2010 campaign.

The rest of the deliverable is organized as follows. We briefly review the evaluation design of the 2010 evaluation campaign (§2), presenting the evaluation data sets, criteria and metrics and the list of participants per track. Then, we illustrate how to run a complete evaluation cycle using the web interface (§3). The preliminary evaluation results are then presented (§4) for each SEALS track. Finally, we comment on the main lessons learned (§5) and conclude the paper (§6).

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

2. Evaluation Campaign

In the following we describe the test data used within the first SEALS campaign and motivate its choice. We continue with a short description of the criteria and metrics used in the context of this campaign. Then we describe the overall methodology we followed to conduct the campaign. Finally, we give an overview on the participants of the campaign.

2.1 Test data

Anatomy, *Benchmark* and *Conference* have been included in the SEALS evaluation modality. The reason for this is twofold: on the one hand these data sets are well known to the organizers and have been used in many evaluations contrary to the test cases of the instance data sets, for instance. On the other hand these data sets come with a high quality reference alignment which allows for computing the compliance based measures, such as precision and recall.

2.1.1 Benchmark test data

The domain of this first test is Bibliographic references. It is based on a subjective view of what must be a bibliographic ontology. There may be many different classifications of publications, for example, based on area and quality. The one chosen here is common among scholars and is based on publication categories; as many ontologies (tests #301-304), it is reminiscent to BibTeX.

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

Simple tests (1xx) such as comparing the reference ontology with itself, with another irrelevant ontology (the wine ontology used in the OWL primer) or the same ontology in its restriction to OWL-Lite;

Systematic tests (2xx) obtained by discarding features from some reference ontology. It aims at evaluating how an algorithm behaves when a particular type of information is lacking. The considered features were:

- *Name of entities* that can be replaced by random strings, synonyms, name with different conventions, strings in another language than English;
- *Comments* that can be suppressed or translated in another language;
- *Specialization hierarchy* that can be suppressed, expanded or flattened;
- *Instances* that can be suppressed;

- *Properties* that can be suppressed or having the restrictions on classes discarded;
- *Classes* that can be expanded, i.e., replaced by several classes or flattened.

Four real-life ontologies of bibliographic references (3xx) found on the web and left mostly untouched (there were added xmlns and xml:base attributes).

Since one goal of these tests is to offer a permanent benchmark to be used by many, the test is an extension of the 2004 EON Ontology Alignment Contest, whose test numbering it (almost) fully preserves.

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

2.1.2 Anatomy test data

The data set of this track has been used since 2007. For a detailed description we refer the reader to the OAEI 2007 [6] results paper. The ontologies of the anatomy track are the NCI Thesaurus describing the human anatomy, published by the National Cancer Institute (NCI), and the Adult Mouse Anatomical Dictionary, which has been developed as part of the Mouse Gene Expression Database project. Both resources are part of the Open Biomedical Ontologies (OBO). The alignment between these ontologies has been created by experts of the domain [1].

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

Subtask #1 The matcher has to be applied with its standard settings.

Subtask #2 An alignment has to be generated that favors precision over recall.

Subtask #3 An alignment has to be generated that favors recall over precision.

Subtask #4 A partial reference alignment has to be used as additional input.

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

The harmonization of the ontologies applied in the process of generating a reference alignment (see [1] and [6]), 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

[1] `http://oaei.ontologymatching.org/2010/benchmarks/`

following analysis. The partial reference alignment used in subtask #4 is the union of all trivial correspondences and 54 non-trivial correspondences.

Due the experiences made in the past, we decided to slightly modify the test data set for the 2010 evaluation. We removed some doubtful subsumption correspondences and added a number of disjointness statement at the top of the hierarchies to increase the expressivity of the data set. Furthermore, we eliminated three incorrect correspondences. The reference alignment is now coherent with respect to the ontologies to be matched.[2]

2.1.3 Conference test data

The collection consists of sixteen ontologies in the domain of organizing conferences. Ontologies have been developed within the OntoFarm project[3]. In contrast to last year's conference data set, this year is supported by the SEALS evaluation service.

The main features of this test set are:

- *Generally understandable domain.* Most ontology engineers are familiar with organizing conferences. Therefore, they can create their own ontologies as well as evaluate the alignment among their concepts with enough erudition.
- *Independence of ontologies.* Ontologies were developed independently and based on different resources, they thus capture the issues in organizing conferences from different points of view and with different terminologies.
- *Relative richness in axioms.* Most ontologies were equipped with DL axioms of various kinds, which opens a way to use semantic matchers.

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

This year, results of participants will be evaluated by four different methods of evaluation: evaluation based on a reference alignment, manual labeling, data mining method, and logical reasoning. In the results section we will report on three of them. Similarly to OAEI 2009, we have still 21 alignments (with some corrections in comparison with the previous year), which correspond to the complete alignment space between 7 ontologies from the data set. Manual evaluation will produce statistics such as precision and will also help in the process of improving and building a reference alignment to be used for further evaluation campaigns.

[2]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`

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

2.2 Evaluation criteria and metrics

The diverse nature of OAEI data sets, specially in terms of the complexity of test cases and presence/absence of (complete) reference alignments, requires to use different evaluation measures. For the three data sets in the SEALS modality, compliance of matcher alignments with respect to the reference alignments is evaluated. In the case of *Conference*, where the reference alignment is available only for a subset of test cases, compliance is measured over this subset. The most relevant measures are precision (true positive/retrieved), recall (true positive/expected) and f–measure (aggregation of precision and recall). These metrics are also partially considered or approximated for the other data sets which are not included in the SEALS modality (standard modality).

For *Conference*, alternative evaluation approaches have been applied, such as manual labeling. These approaches require a more deep analysis from experts than traditional compliance measures. For the first version of the evaluation service, we concentrate on the most important compliance based measures because they do not require a complementary step of analyse/interpretation from experts, which is mostly performed manually and outside an automatic evaluation cycle. However, such approaches will be progressively integrated into the SEALS infrastructure.

Nevertheless, for 2010, the generated alignments are stored in the results database and can be retrieved by the organizers easily. It is thus still possible to exploit alternative evaluation techniques subsequently, as it has been done in the previous OAEI campaigns. In particular, we will compute the degree of coherence for the *Conference* alignments stored as results. We have already developed an algorithm for that purpose, however, further experiments – conducted semi-automatically in the context of this years evaluation campaign – are required before it can be deployed in the SEALS platform as stable component.

All the criteria above are about alignment quality. A useful comparison between systems also includes their efficiency, in terms of runtime and memory consumption. The best way to measure efficiency is to run all systems under the same controlled evaluation environment. In previous OAEI campaigns, participants have been asked to run their systems on their own and to inform about the elapsed time for performing the matching task. Using the web based evaluation service, runtime cannot be correctly measured due the fact that the systems run in different execution environments and, as they are exposed as web services, there are potential network delays.

2.3 General methodology

The process of the complete evaluation campaign can divided into four phases. In the preparatory phase data sets are prepared and provided to participants. In the phase of preliminary testing participants ensure that they can work with the data regarding format and technical issues. In the execution phase participants use their algorithms to automatically match the ontologies, and in the evaluation phase the alignments provided by the participants are evaluated and compared.

2.3.1 Preparatory phase

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

2.3.2 Preliminary tests

In this phase, participants were invited to test their systems in order to ensure that the systems can load the ontologies to be aligned and generate the alignment in the correct format, the Alignment format expressed in RDF/XML [4]. Participants have been requested to provide (preliminary) results by August 30th.

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

2.3.3 Execution phase

During the execution phase, participants used their systems to automatically match the ontologies from the test cases. Participants have been asked to use one algorithm and the same set of parameters for all tests in all tracks. It is fair to select the set of parameters that provide the best results (for the tests where results are known). Beside parameters, the input of the algorithms must be the two ontologies to be matched and any general purpose resource available to everyone, i.e., no resource especially designed for the test. In particular, participants should not use the data (ontologies and reference alignments) from other test cases to help their algorithms. In most cases, ontologies are described in OWL-DL and serialized in the RDF/XML format.

For the standard OAEI modalities, participants had to run their systems on their own machines and submit the results via mail to the organizers. SEALS participants ran their systems via the SEALS evaluation service. They got a direct feedback on the results and could validate them as final results. Furthermore, SEALS participants were invited to register their tools by that time in the SEALS portal[4].

Participants also provided the papers that are published hereafter and a link to their systems and their configuration parameters. These paper comprise a system description and an analysis of the results from the perspective of the tool developer. In the past it has often been the case that the organizers could not compute and present all required results in time. As consequence, tool developers could not explain or discuss these missing results in their papers. Withe the direct feedback cycle supported by the SEALS service we have solved this problem.

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

2.3.4 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 4th, 2010. In the case of blind tests, only the organizers did the evaluation with regard to the withheld reference alignments.

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

The standard evaluation measures are precision and recall computed against the reference alignments. For the matter of aggregation of the measures, we use weighted harmonic means (weights being the size of the true positives). This clearly helps in the case of empty alignments. Another technique that has been used is the computation of precision/recall graphs so it was advised that participants provide their results with a weight to each correspondence they found. New measures addressing some limitations of precision and recall have also been used for testing purposes as well as measures compensating for the lack of complete reference alignments.

2.4 Participants

The OAEI campaign had 15 participants in 2010 [5]. Regarding the SEALS tracks, 11 participants have registered their results for Benchmark, 9 for Anatomy and 8 for Conference. Some participants in Benchmark have not participated in Anatomy or Conference and vice-versa. The list of participants is summarized in Table 2. In this table, confidence stands for the type of result returned by a system: it is ticked when the confidence has been measured as non boolean value.

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

Table 2.1: Participants and the state of their submissions.

Regarding previous OAEI compaings, since a few years, the number of participating systems has remained roughly stable: 4 participants in 2004, 7 in 2005, 10 in 2006, 17 in 2007, 13 in 2008, 16 in 2009 and 15 in 2010. The number of covered runs has decreased more than expected: 37 in 2010, 53 in 2009, 50 in 2008, and 48 in 2007.

This may be due to the increasing specialization of tests: some systems are specifically designed for instance matching or for anatomy.

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

3. Running and Registering an Evaluation

Tool developers can use the evaluation service for testing the versions and configurations of their systems and/or for participating in a campaign. In the context of a campaign, they can select the most suitable results from previous testing and register them for the campaign. The evaluation service offers functionalities for both contexts, as we will illustrate in the following.

For illustrating a complete evaluation cycle, we have extended the Anchor-Flood system [12] with the web service interface[1]. This system has participated in the two previous OAEI campaigns and is thus a typical evaluation target. The current version of the web application described in the following is available at `http://seals.inrialpes.fr/platform/`.

In order to start an evaluation, one must specify the URL of the matcher service, the class implementing the required interface and the name of the matching system to be evaluated (Figure 3.1). Three of the OAEI data sets have been selected, namely Anatomy, Benchmark and Conference. In this example, we have used the conference test case.

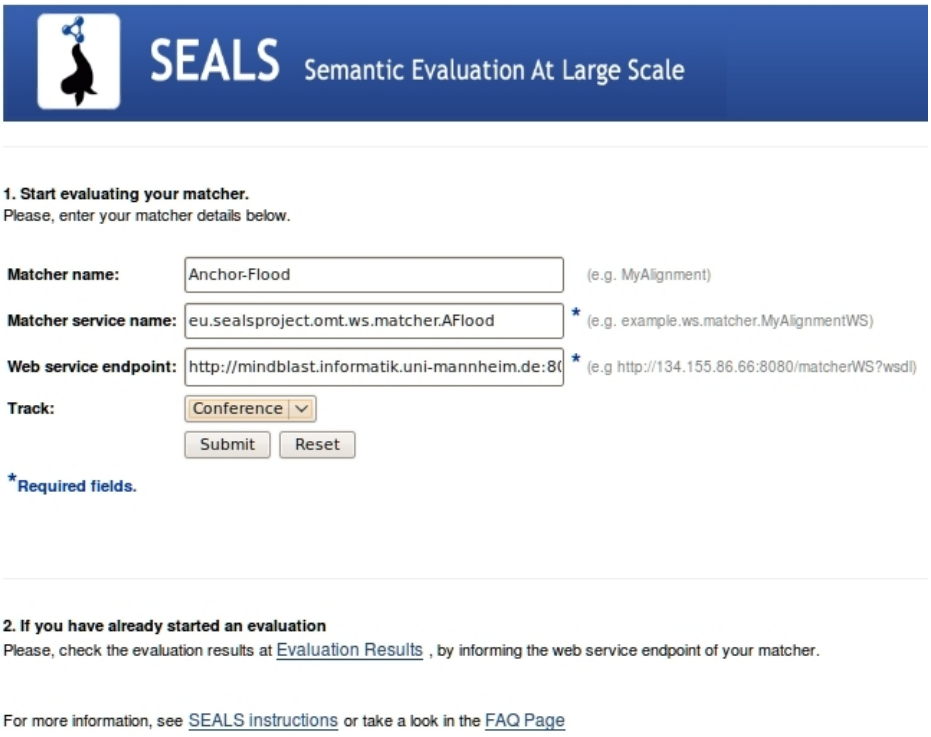


Figure 3.1: Specifying a matcher endpoint as evaluation target.

Submitting the form data, the BPEL workflow is invoked. It first validates the specified web service as well as its output format. In case of a problem, the concrete validation error is displayed to the user as direct feedback (Figure 3.2). In case of

[1] Available at `http://mindblast.informatik.uni-mannheim.de:8080/sealstools/aflood/matcherWS?wsdl`

a successfully completed validation, the system returns a confirmation message and continues with the evaluation process.

Figure 3.2: Example of error message displayed to the user.

Every time an evaluation is conducted, results are stored under the endpoint address of the deployed matcher (Figure 3.3).

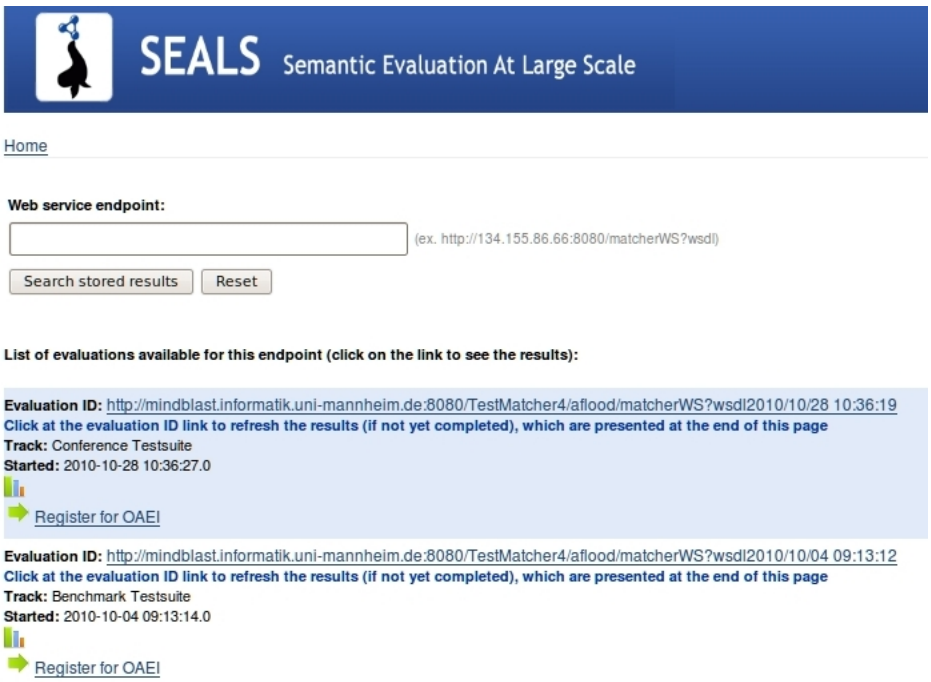


Figure 3.3: Listing of available evaluation results for the specific endpoint.

The results are displayed as a table (Figure 3.4), when clicking on one of the three evaluation IDs in Figure 3.3. The results table is (partially) available while the

Test	Precision	Recall	F-Measure	Status	Alignment
cmt-cocus				No reference alignment	
cmt-conference	0.300	0.375	0.333	completed	
cmt-confious				No reference alignment	
cmt-confof	0.455	0.313	0.370	completed	
cmt-crs_dr				No reference alignment	
cmt-edas				not completed	
cmt-ekaw				not completed	
cmt-iasted				not completed	
cmt-linklings				not completed	

Figure 3.4: Display results of an evaluation.

Correct correspondences (generated and in the reference alignment)

Person = Person
Conference = Conference
Review = Review
PaperAbstract = Abstract
email = has_an_email
ProgramCommittee = Program_committee

Back to top

Incorrect correspondences (generated but not in the reference alignment)

Chairman = Co-chair
endReview = has_a_review
Paper = Paper
ProgramCommitteeChair = Reviewed_contribution
writeReview = reviews

Back to top

Missing correspondences (not generated but in the reference alignment)

assignExternalReviewer = invites_co-reviewers
SubjectArea = Topic
PaperAbstract = Extended_abstract
assignedByReviewer = invited_by

Figure 3.5: View on an alignment.

evaluation itself is still running. By reloading the page from time to time, users can see the progress of an evaluation that is still running. In the results table, for each test case, precision and recall are listed. Moreover, a detailed view on the alignment results is available (Figure 3.5), when clicking on the alignment icon in Figure 3.4. This detailed view lists those correspondences, that (a) have been generated and are in the reference alignment (true positives), that (b) have been generated but are not in the reference alignment (false positives), and that (c) have not been generated but are in the reference alignment (false negatives).

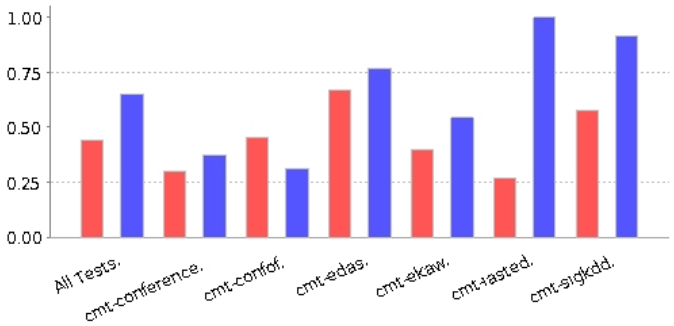
Furthermore, the user can visualize the results in an OLAP application (Figure 3.6), by clicking on the plot figure in Figure 3.3.

For registering a set of results for the campaign, the user selects the evaluation experiment he wants to register ("Register for OAEI" in Figure 3.3). This process must be done for each track. The user, then, informs the official name of the tool and its contact mail (Figure 3.7).

On the other hand, organizers have a tool for accessing the results registered for

	Measures	
Test	Precision	Recall
All Tests	.44	.65
cmt-conference	.30	.38
cmt-confof	.45	.31
cmt-edas	.67	.77
cmt-ekaw	.40	.55
cmt-iasted	.27	1.00
cmt-sigkdd	.58	.92

Slicer: [ID=http://mindblast.informatik.uni-mannheim.de:8080/TestMatcher4/aflood/



Slicer: ID=http://mindblast.informatik.uni-mannheim.de:
8080/TestMatcher4/aflood/matcherWS?wsdl2010/10/28 10:36:19
Precision. Recall.

Figure 3.6: Using OLAP for results visualization.

Figure 3.7: Registering the results for the campaign.

the campaign (Figures 3.8 and 3.9) as well as all evaluations being carried out in the evaluation service, even the evaluation executed for testing purposes, in a similar list from what is illustrated in Figure 3.9).

Organizers can download the set of alignment for each track ("Download all evaluation per track" in Figure 3.9). This option is specially useful in the case of the

SEALS Semantic Evaluation At Large Scale

Home | Logout

List of tools registered for the 2010 evaluation campaing:

Tool	Contact	Tracks
AFlood	christian@informatik.uni-mannheim.de	Anatomy Testsuite Benchmark Testsuite Conference Testsuite
AgrMaker	ifc@cs.uic.edu	Anatomy Testsuite Benchmark Testsuite Conference Testsuite
AROMA	jerome.david@inrialpes.fr	Anatomy Testsuite Benchmark Testsuite Conference Testsuite
ASMOV	reggie@infotechsoft.com	Anatomy Testsuite Benchmark Testsuite Conference Testsuite
BLOOMS	cpesquita@xldb.di.fc.ul.pt	Anatomy Testsuite
CODI	jan@informatik.uni-mannheim.de	Anatomy Testsuite Benchmark Testsuite Conference Testsuite
Ef2Match	watsonchua@pmail.ntu.edu.sg	Anatomy Testsuite Benchmark Testsuite Conference Testsuite
Falcon	whu1982@gmail.com	Benchmark Testsuite Conference Testsuite
GeRMeSMB	quix@dbis.rwth-aachen.de	Anatomy Testsuite Benchmark Testsuite Conference Testsuite
MapPSO	bock@fzi.de	Benchmark Testsuite
NBJLM	jackws66@yahoo.com	Anatomy Testsuite
RiMOM	zcwang@keg.cs.tsinghua.edu.cn	Benchmark Testsuite
SOBOM	peigang.xu@gmail.com	Anatomy Testsuite Benchmark Testsuite Conference Testsuite
TaxoMap	hamdi@lri.fr	Anatomy Testsuite Benchmark Testsuite

Figure 3.8: List of all registered tools for the campaign. Organizers can see the results for that tool by clicking at its name link.

Conference track, where organizers apply alternative approaches, not yet supported by the evaluation service, for evaluating the generated alignments.

Figure 3.9: List of all registered results for all SEALS tracks.

4. Evaluation Results

In the following we present the preliminary results of the campaign. This chapter is divided in three sections, in which we describe the results for Benchmark, Anatomy, and Conference dataset.

4.1 Benchmark results

Eleven systems have participated in the benchmark track of this year's campaign (see Table 2.1). 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 4.1 shows the results, by groups of tests (means of results corresponding to harmonic means). The symmetric relaxed measure corresponds to the relaxed precision and recall measures of [3]. For comparative purposes, the results of systems that have participated last year are also provided. We display the results of participants as well as those given by some simple edit distance algorithm on labels (edna). The computed values are real precision and recall and not an average of precision and recall. The full results are on the OAEI web site.

As shown in Table 4.1, two systems are ahead: ASMOV and RiMOM, with AgrMaker as close follower, while SOBOM, GeRMeSMB and Ef2Match, respectively, had presented intermediary values of precision and recall. In the 2009 campaign, Lily and ASMOV were ahead, with aflood and RiMOM as followers, while GeRoME, AROMA, DSSim and AgrMaker had intermediary performance. The same group of best matchers has been presented in both campaigns. No system had strictly lower performance than edna.

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

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

As last year, the apparently best algorithms provide their results with confidence measures. It is thus possible to draw precision/recall graphs in order to compare them.

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

Table 4.1: Benchmark results.

Figure 4.1 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 2.1). 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.

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

The results presented in Table 4.1 and those displayed in Figure 4.1 single out the same group of systems, ASMOV, RiMOM and AgrMaker, which seem to perform these tests at the highest level. Of these, ASMOV has slightly better results than the two others. So, this confirms the observations on raw results.

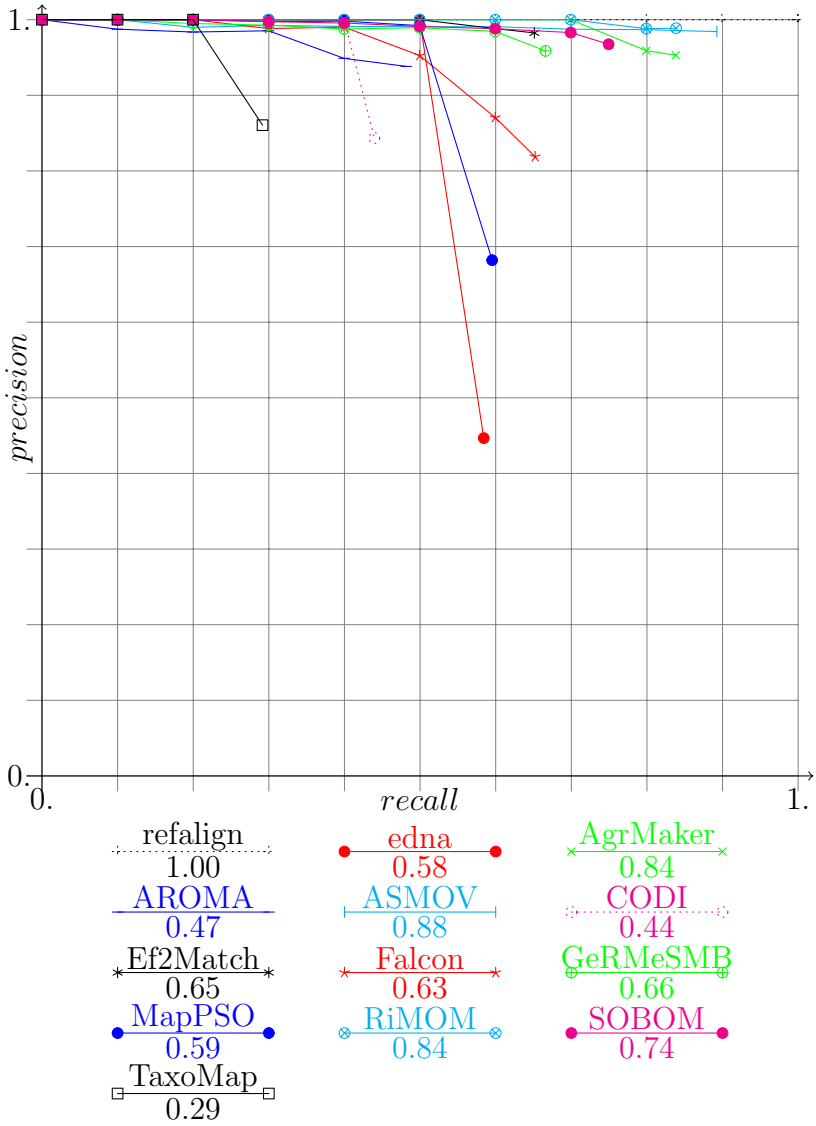


Figure 4.1: Precision/recall graphs for benchmarks.

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

4.2 Anatomy

In 2010 nine systems participated in the Anatomy track. While the number of participants is nearly stable over the last four years of the OAEI, we find in 2010 more systems that participated for the first time (5 systems compared to 2 systems in the years before). See Table 4.2 for an overview. Four of the newcomers participate also in other tracks, while NBJLM participates only in the Anatomy track. NBJLM is thus together with AgreementMaker (AgrMaker) a system that uses a track-specific parameter setting. Taking part in several tracks with a standard setting makes it obviously much harder to obtain good results in a specific track. Notice that in the current version of the SEALS service the configuration of a tool is not under control of the organizers. This problem will be solved when the tools are executed on the SEALS platform in the next year.

In the last row of Table 4.2, the average of F-measures per year in subtask #1 is shown. We observe significant improvements over time. However, the measured improvements decrease over time and seem to reach a top (2007 +12% → 2008 +5% → 2009 +2% → 2010). We have marked the participants with an F-measure ≥ 0.8 with a + symbol. Note that in each of the previous years, only two systems reached this level, while in 2010 six systems reached a higher value than 0.8. This has – at least partially – been caused by the possibility to run evaluations with the testset, which has been used extensively by some of the particpants.

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

Table 4.3: Results for subtasks #1, #2 and #3 in terms of precision, recall (in addition recall+ for #1 and #3) and F-measure.

In the previous years, OAEI organizers reported about runtimes that have been measured by the participants. The differences we observed – from several minutes to several days – could not be explained by the use of different hardware. However, these differences became less significant over the years and in 2009 all systems except one required between 2 and 30 minutes. Therefore, we abstained from an analysis of runtimes this year. In 2011, we plan to execute the matching systems on the SEALS platform to enable an exact measurement of runtimes not biased by differences in hardware equipment.

The results for subtask #1 are presented in Table 4.3 ordered with respect to the achieved F-measure. Systems marked with a * do not participate in other tracks or have chosen a setting specific to this track. Note that ASMOV modified its standard settting in a very restricted way (activating UMLS as additional resource). Thus, we did not mark this system.

In 2010, AgreementMaker (AgrMaker) generates the best alignment with respect to F-measure. Moreover, this result is based on a high recall compared to the systems on the following positions. This is a remarkable result, because even the SAMBO system of 2007 could not generate a higher recall with the use of UMLS. However, we have to mention again that AgreementMaker uses a specific setting for the anatomy track. AgreementMaker is followed by three participants (Ef2Match, NBJLM and SOBOM) that share a very similar characteristic regarding F-measure and observed precision score. All of these systems clearly favor precision over recall.

In the following we use the recall+ measure as defined in [6] for further analysis. It measures how many non trivial correct correspondences, not detectable by string equivalence, can be found in an alignment. The top three systems with respect to recall+ regarding subtask #1 are AgreementMaker, NBJLM and ASMOV. Only ASMOV has participated in several tracks with the same setting. Obviously, it is not easy to find a large amount of non-trivial correspondences with a standard setting. In 2010, five system participated in subtask #3. The top three systems regarding recall+ in this task are GeRoMe-SMB (GeRMeSMB), AgreementMaker and ASMOV. Since a specific instruction about the balance between precision and recall is missing in the description of the task, the results vary to a large degree. GeRoMe-SMB detected

System	Δ-Precision	Δ-Recall	Δ-F-measure
AgrMaker	$+0.025_{\ 0.904\rightarrow0.929}$	$-0.025_{\ 0.876\rightarrow0.851}$	$-0.002_{\ 0.890\rightarrow0.888}$
ASMOV	$+0.029_{\ 0.808\rightarrow0.837}$	$-0.016_{\ 0.824\rightarrow0.808}$	$+0.006_{\ 0.816\rightarrow0.822}$
CODI	$-0.002_{\ 0.970\rightarrow0.968}$	$+0.030_{\ 0.716\rightarrow0.746}$	$+0.019_{\ 0.824\rightarrow0.843}$
SAMBOdtf_{2008}	$+0.021_{\ 0.837\rightarrow0.856}$	$+0.003_{\ 0.867\rightarrow0.870}$	$+0.011_{\ 0.852\rightarrow0.863}$

Table 4.4: Changes in precision, recall and F-measure based on comparing $A_1 \cup R_p$ and A_4 against reference alignment R.

83.8% of the correspondences marked as non-trival at a low precision of 8%. AgreementMaker and ASMOV modified their settings only slightly, however, they were still able to detect 70% and 53.8% of all non trivial correspondences.

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

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.[1] 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.

Results are presented in Table 4.4. 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 tendency we observed in the past years (compare with the results for SAMBOdtf in 2008). The effects of this strategy are not very strong. However, as argued in the previous years, the input alignment has a characterictis that makes it hard to exploit this information.

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

[1] 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.

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

Table 4.5: Confidence threshold, precision and recall for optimal F-measure for each matcher.

The availability of the SEALS evaluation service surely had an effect on the results submitted in 2010. In the future, we plan to extend the data set of the anatomy track with additional ontologies and reference alignments to a more comprehensive and general track covering different types of biomedical ontologies. In particular, we will not publish the complete set of reference alignments to conduct a part of the evaluation experiment in the blind mode. This requires, however, to find and analyze interesting and well-suited data sets.

4.3 Conference

The conference test set introduces matching several more-or-less expressive ontologies. Within this track the results of participants will be evaluated using diverse evaluation methods. Note that only the compliance based measures of precision, recall and F-measure are so far supported by the SEALS evaluation service, while the remaining methods are applied semi-automatically by the OAEI organizers. For that purpose the raw results stored in the SEALS platform have been made available to the organizers.

Eight systems participanted to the conference track: AgreementMaker (AgrMaker), AROMA, ASMOV, CODI, Ef2Match, Falcon, GeRMeSMB and SOBOM. Contrary to the previous years, all participants delivered results for all 120 testcases of the dataset. This might be caused by the evaluation workflow that forces all systems to apply their matcher to each of the testcases. Note that a failure caused by the matching tool is highlighted explictly by the evaluation service. As a result systems might have become more robust.

We evaluated the results of participants against a reference alignment. It includes all pairwise combinations of a subset of seven ontologies (i.e. 21 alignments). For a better comparison, we established the confidence threshold which provides the highest average F-measure (Table 4.5). Precision, Recall, and F-measure are given for this optimal confidence threshold. The dependency of F-measure on confidence threshold can be seen from Figure 4.2. 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. Ef2Match and

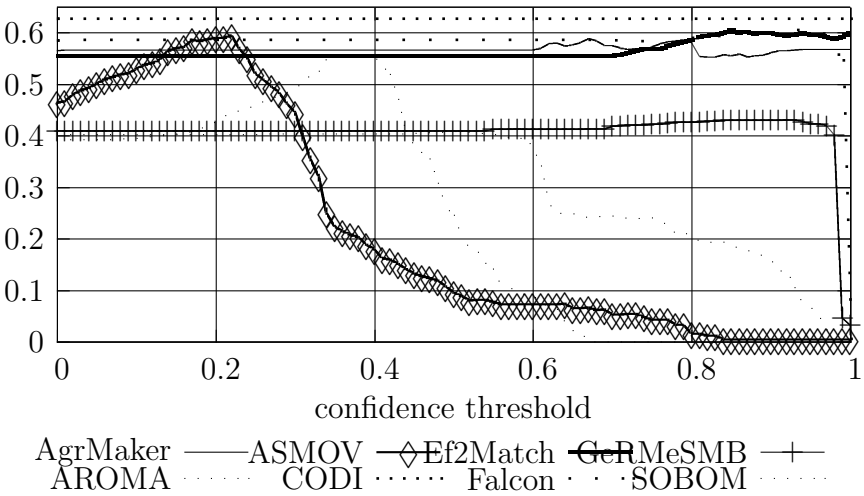


Figure 4.2: F-measures depending on confidence.

ASMOV with 60% of F-Measure). However, we should take into account that this evaluation has been made over a subset of all alignments (one fifth).

We compared the performance of participants wrt. last two years (2008, 2009). There are three matchers which also participated in 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 58% in 2010 regarding highest average F-measure. Finally, AROMA participated with the same highest average F-measure in both years, 2009 and 2010.

Since reference alignments are only available for a subset of all possible combinations, we conducted additionally an evaluation based on posterior manual labeling of a sample. This year we take the most secure, i.e., with highest confidence, correct correspondences as a population for each matcher. Particularly, we evaluate 100 correspondences per matcher randomly chosen from all correspondences of all 120 alignments with confidence 1.0 (sampling). Because AROMA, ASMOV, Falcon, GeRMeSMB and SOBOM do not have enough correspondences with 1.0 confidence we take 100 correspondences with highest confidence.

In Table 4.6 you can see approximated precision values for each matcher over its population of best correspondences. N is a population of all the best correspondences for one matcher. n is a number of randomly chosen correspondences so it is 100 best correspondences for each matcher. TP is a number of correct correspondences from the sample, and P* is an approximation of precision for the correspondences in each population; additionally there is a margin of error computed as: $\frac{\sqrt{(N/n)-1}}{\sqrt{N}}$ based on [14]. From Table 4.6 we can conclude that CODI, Falcon and AgreementMaker have the best precision (higher than 90%) over their 100 more confident correspondences.

In 2008 the OAEI organizers evaluated for the first time the coherence of the submitted alignments. Alignment coherence can be computed automatically, contrary

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

Table 4.6: Approximated precision for 100 best correspondences for each matcher.

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 4.7: Degree of incoherence and size of alignment in average for the optimal a-posteriori threshold.

to the labeling technique described above. For the current campaign we have abstained from an integration in the SEALS platform. However, we have gained more experiences and are now prepared to include a corresponding evaluation service in the SEALS platform for the next evaluation campaign.

We already discussed the Maximum Cardinality measure – first proposed in [9] – in [2]. Since the underlying problem is the NP-complete problem of computing a hitting-set, there exist no algorithm that can solve the problem in acceptable time for larger problem instances. For that reason we have used a timeout of 1000 seconds. In case the search algorithm does not find a optimal solution in time, a lower bound for the degree of incoherence is returned. This occurred for less the 5% of all analyzed alignments. The reasoning required to compute the degree of incoherence has been performed with the Pellet reasoner [13].

The results are depicted in Table 4.7. The prefix > is added whenever the search algorithm stopped in a testcase due to the timeout prior to finding the solution. The actual value can be expected to be slightly higher. The table shows the average for all testcases of the conference track except the testcases where the ontologies confious and linklings are involved.[2] Note that we did not use the original alignments, but the alignments with optimal threshold as explained above. Nevertheless the average size of the resulting alignment in number of correspondences N varies still to a large degree. From all participants CODI [11] generated the alignment with lowest degree of incoherence. Note that this result is also to a large degree caused by the small size of alignments where the occurrence of an incoherence is less probable. Taking into account the size of the alignment, the ASMOV [8] system generates a remarkable result. Even though the alignments of ASMOV comprise the highest number of correspondences, the degree of incoherence 5.6% is relatively small.

[2]These ontologies resulted in some combinations of ontologies and alignments in reasoning problems. This issue has to be analyzed in depth to provide a stable evaluation service applicable to all testcases.

5. Lessons Learned

The new technology introduced in the OAEI affected both tool developers and organizers to a large degree. In the following we highlight some of the outcomes and describe the lessons we learned from the experiences made so far.

OAEI Continuity We were not sure that switching to an automated evaluation would preserve the success of OAEI, given that the effort of implementing a web service interface was required from participants. This has been the case. The number of participants is similar to the numbers we observed in the last years. Moreover, the conference track has more participants than in the last years. This might be caused by the fact that many participants mainly focus on participating in the benchmark track. However, once the interface is implemented it is just one additional mouse click to participate also in the conference track.

Implementing the interface As already argued, implementing the web service interface requires some effort on side of a tool developer. We stayed in contact with some of the tool developers during this process. Thereby, we observed that the time required for implementing the interface varied between several hours and several days depending on the technical skills of each developer. We also became aware that the first version of the provided tutorial contained some unclear information resulting in problems for some participants. From the feedback of the developers, we have improved the tutorial. Another typical problem is related to the fact that some tool developer had only restricted access to a machine that is available from the Internet. These problems could finally be solved, however, system administrators of the particular company or research institute should be contacted early by participants.

Usage by participants Once technical problems had been solved, the evaluation service has been used by some of the participants in the phase of preliminary testing extensively. Obviously, the direct feedback of the evaluation service has supported the process of a formative evaluation well. Other participants used the service only for submitting their final results. This might have been caused partially by a suboptimal runtime performance during the first weeks. Even though we finally solved the underlying problems, these problems might have been the reason for some participants to abandon from the use during the first weeks. Once the problems have been solved, we contacted each participant in order to explain the problems and many of them started to use the system again.

Organizational effort On side of the organizers, the evaluation service reduced the effort of checking the formal correctness of the results to a large degree. In the past, it was required to communicate many of the problems in a time-consuming multilevel process. Typical examples are invalid xml, missing or incorrect namespace information, unsupported types of relations in generated alignments, incorrect directory structure and an incorrect naming style used for the submissions. All of these problems are now directly forwarded to the tool developer in an error message or in a preliminary

result interpretation that does not fit with the expectations. Moreover, the organizers could analyze the results so far submitted at any time and had an overview on the participants using the system.

Evaluation and analysis services While some analysis methods are already available, a number of specific services and operations is still missing. The graphical support of the OLAP visualization does, for example, not support the generation of precision and recall graphs frequently used by OAEI organizers. In particular, evaluation and visualization methods specific for ontology matching are not supported. However, most of these operations are already implemented in the Alignment API and will be made available in the future. On the other hand we already developed and tested a component for measuring the degree of incoherence of an alignment. We will try to include this additional metric in the next version of the SEALS platform.

Automatic test generation The benchmark test case is not discriminant enough between systems. The results presented above have shown this. Next year, we plan to introduce controlled automatic test generation in the SEALS platform. This will improve the situation and allow OAEI organizers and SEALS campaign organizers to construct testsuites with the required type of testcases.

Configuration and control over matching systems We have seen that not all systems followed the general rule to use the same set of parameters in all tracks. This problem will be solved when we deploy the systems in the SEALS platform in the following campaign. However, we also have to support different system configurations in a controlled environment. To run a system with specific settings is an explicit requirement of subtask #2 and #3 of the anatomy track. It should thus be possible to run a system deployed on the SEALS platform with different parameter setting.

6. Final Remarks

This deliverable has presented the results of the 2010 OAEI/SEALS integrated campaign. A subset of the OAEI tracks has been selected to be included in the SEALS modality. The main innovation for the OAEI community was the use of the evaluation service, from which participants have launched their own experiments. This new technology introduced in the OAEI affected both tool developers and organizers to a large degree and has been accepted positively on both sides.

There are several plans for the next campaign. In the current setting, runtime and memory consumption cannot be correctly measured because a controlled execution environment is missing. The same holds for the reproducibility of the results. This is definitely the main issue that we have to approach. Further versions of the SEALS evaluation service will include the deployment of tools in such a controlled environment. We also plan to integrate additional metrics and visualization components as noted above. Furthermore, we try to find more well suited data sets to be used as test suites in the platform and finally we have to develop a test generator that allows a controlled automatic test generation of high quality data sets.

References

[1] Oliver Bodenreider, Terry F. Hayamizu, Martin Ringwald, Sherri De Coronado, and Songmao Zhang. Of mice and men: Aligning mouse and human anatomies. In *AIMA 2005 Symposium Proceedings*, pages 61–65, 2005.

[2] Cássia Trojahn dos Santos, Jérôme Euzenat, Christian Meilicke, and Heiner Stuckenschmidt. D12.1 Evaluation Design and Collection of Test Data for Matching Tools. Technical report, SEALS Project `<http://sealsproject.eu>`, November 2009.

[3] Marc Ehrig, Steffen Staab, and York Sure. Bootstrapping ontology alignment methods with APFEL. In *Proc. 4th International Semantic Web Conference (ISWC)*, volume 3729 of *Lecture notes in computer science*, pages 186–200, Galway (IE), 2005.

[4] Jérôme Euzenat. An API for ontology alignment. In *Proceedings of the 3rd International Semantic Web Conference (ISWC)*, pages 698–712, Hiroshima (JP), 2004.

[5] Jérôme Euzenat, Alfio Ferrara, Christian Meilicke, 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, Natasha Noy, and Arnon Rosenthal, editors, *Proc. 5th ISWC workshop on ontology matching (OM), Shanghai (Chine)*, pages 1–35, 2010.

[6] Jérôme Euzenat, Antoine Isaac, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Svab, Vojtech Svatek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative 2007. In Pavel Shvaiko, Jrme Euzenat, Fausto Giunchiglia, and Bin He, editors, *Proceedings of the 2nd ISWC international workshop on Ontology Matching, Busan (KR)*, pages 96–132, 2007.

[7] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer, Heidelberg (DE), 2007.

[8] Yves R. Jean-Mary, E. Patrick Shironoshitaa, and Mansur R. Kabuka. Ontology matching with semantic verification. *Web Semantics: Science, Services and Agents on theWorldWideWeb*, 158, 2009.

[9] Christian Meilicke and Heiner Stuckenschmidt. Incoherence as a basis for measuring the quality of ontology mappings. In *Proc. of the ISWC 2008 Workshop on Ontology Matching*, Karlsruhe, Germany, 2008.

[10] Christian Meilicke, Cassia Trojahn, Jérôme Euzenat, and Heiner Stuckenschmidt. Services for the automatic evaluation of matching tools. Technical Report D12.2, SEALS Project, July 2010.

[11] Jan Noessner. CODI: Combinatorial optimization for data integration: results for OAEI 2010. In *Proceedings of the ISWC 2010 Workshop on Ontology Matching*, Shanghai, China, 2010.

[12] Hanif Seddiqui and Masaki Aono. Anchor-flood: results for OAEI 2009. In *Proceedings of the ISWC 2009 workshop on ontology matching*, Washington DC, USA, 2009.

[13] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: a practical OWL-DL reasoner,. *Journal of Web Semantics*, Volume 5, Issue 2:51–53, 2007.

[14] Willem Robert van Hage, Antoine Isaac, and Zharko Aleksovski. Sample evaluation of ontology-matching systems. In *Proc. 5th International Workshop on Evaluation of Ontologies and Ontology-based Tools (EON 2007), collocated with ISWC-2007*, pages 41–50, Busan (Korea), 2007.