Ontology Alignment Evaluation Initiative: six years of experience

Jérôme Euzenat[1], Christian Meilicke[2], Heiner Stuckenschmidt[2], Pavel Shvaiko[3], and Cássia Trojahn[1]

[1] INRIA & LIG, Grenoble, France
{jerome.euzenat, cassia.trojahn}@inria.fr
[2] University of Mannheim, Germany
{christian, heiner}@informatik.uni-mannheim.de
[3] Informatica Trentina S.p.A., Trento, Italy
pavel.shvaiko@infotn.it

Abstract. In the area of semantic technologies, benchmarking and systematic evaluation is not yet as established as in other areas of computer science, e.g., information retrieval. In spite of successful attempts, more effort and experience are required in order to achieve such a level of maturity. In this paper, we report results and lessons learned from the Ontology Alignment Evaluation Initiative (OAEI), a benchmarking initiative for ontology matching. The goal of this work is twofold: on the one hand, we document the state of the art in evaluating ontology matching methods and provide potential participants of the initiative with a better understanding of the design and the underlying principles of the OAEI campaigns. On the other hand, we report experiences gained in this particular area of semantic technologies to potential developers of benchmarking for other kinds of systems. For this purpose, we describe the evaluation design used in the OAEI campaigns in terms of datasets, evaluation criteria and workflows, provide a global view on the results of the campaigns carried out from 2005 to 2010 and discuss upcoming trends, both specific to ontology matching and generally relevant for the evaluation of semantic technologies. Finally, we argue that there is a need for a further automation of benchmarking to shorten the feedback cycle for tool developers.

Keywords: Evaluation, experimentation, benchmarking, ontology matching, ontology alignment, schema matching, semantic technologies.

1 Introduction

The past ten years have witnessed impressive development in the area of semantic technologies, mostly driven by the idea of creating a semantic web [4] as a source of information that is accessible by machines. This development has been enabled by the standardization of representation languages for knowledge on the web, in particular RDF and OWL. Based on these languages, many tools have been developed to perform various tasks on the semantic web, such as searching, querying, integrating and reasoning about semi-structured information. Standards were an important factor for the development of software tools supporting semantic web applications. However, a crucial step in their large scale adoption in real world applications will be the ability to

determine the quality of a system in terms of its expected performance on realistic data. This means that systematic evaluation of semantic technologies is an important topic.

A major and long term goal of evaluation is to help developers of such systems to improve them and to help users evaluating the suitability of the proposed systems to their needs. The evaluation should thus be run over several years in order to allow for adequate measurement of the evolution of the field. Evaluation should also help assessing absolute results, i.e., what are the properties achieved by a system, and relative results, i.e., how these results compare to the results of other systems.

One particular kind of evaluation is benchmarking. A benchmark is a well-defined set of tests on which the results of a system or a subsystem can be measured [9]. It should enable to measure the degree of achievement of proposed tasks on a well-defined scale (that can be achieved or not). It should be reproducible and stable, so that it can be used repeatedly for: (i) testing the improvement or degradation of a system with certainty and (ii) situating a system among others. A medium term goal for evaluation efforts is to set up a collection of reference sets of tests, or benchmark suites for assessing the strengths and weaknesses of the available tools and to compare their evolution with regard to these references. Building benchmark suites is valuable not just for groups of people who participate in planned evaluations but for all the community, since system designers can make use of them at any time and compare their results with those of the other systems.

In this paper, we focus on the Ontology Alignment Evaluation Initiative (OAEI)[4] which carries out annual campaigns for the evaluation of ontology matching tools. Ontology matching is an important functionality in many applications as it is the basis for linking information, e.g., from heterogeneous sources into a common model that can be queried and reasoned upon. Initially, the focus of OAEI was on the task of matching different ontologies rather than on the data itself. More recently, however, the focus is being extended to include data matching algorithms as well. 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 OAEI ambition is that from such evaluations, tool developers can learn and improve their systems, thus extending the state of the art in ontology matching.

The goal of this paper is to present the state of the art in evaluating ontology matching. For this purpose, we draw lessons from the six first years of carrying out OAEI focusing on trends we have observed and implications for the further improvement of the OAEI campaigns and the evaluation of semantic technologies in general. Annual OAEI reports [28; 26; 25; 8; 23; 24] present the individual datasets and the results of the different campaigns in detail. In this paper, we take a global view on outcomes of the evaluation campaigns over the years and identify interesting developments, fundamental decisions as well as solved and open problems. Thus, the contributions of the paper are:

- A comprehensive overview of the six years of ontology matching benchmarking in the context of the OAEI initiative accompanied with a rationale for the choice of the datasets used;

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

– The identification and discussion of problems in designing experiments for evaluating matching technologies;
– An analysis of the development of the field of ontology matching on the basis of the results obtained in the different evaluation campaigns;
– Current trends and future challenges of ontology matching evaluation based on our observations and experiences from the OAEI campaigns.

In a nutshell, the lessons learned from the evaluation campaigns can be summarized as follows:

– Systematic ontology matching evaluation indeed allows for measuring the progress of the field in terms of participation to the evaluation campaigns, quality of the matching results and runtime performance;
– It is necessary to be reactive to propose improvements in data sets and evaluation modalities in order to keep or increase the interest in the field;
– Automation is prone to improve the situation on many fronts of ontology matching evaluation, including scalability, variability, and hardness of tests.

The remainder of the paper is structured as follows. In Section 2, we provide an overview of the related work. In Section 3, we introduce the ontology matching problem. Section 4 addresses the problem of designing evaluations for the ontology matching problem and provides some guidelines for the design of future evaluations. Results of the different evaluation campaigns are discussed in Section 5. We first provide background on OAEI, its organization and its development over the years. Then we focus on the progress that has been achieved and how it was measured. In Sections 6 and 7, we summarize our experiences and discuss implications for future evaluation campaigns.

2 Related work on evaluations

Currently, the systematic evaluation of semantic technologies in general still falls behind other fields, such as theorem proving and information retrieval, where benchmarking against standardized datasets is a common practice. Standardized evaluations also provide the basis for a fair comparison of systems according to scientific standards and make it harder to tune results in favor of one or another system. Evaluation initiatives like TPTP (Thousand Problems in Theorem Proving) or TREC (Text Retrieval Conference) that have been carried out on a regular basis for many years have shown that besides the practical benefits of supporting the uptake of technology, systematic and continuous evaluations also lead to a continuous improvement of the field because fundamental problems are better understood and can be addressed more efficiently due to the direct feedback from the frequent evaluation campaigns.

OAEI, presented in this paper, took inspiration from TREC. Indeed, ontology matching is closer to information retrieval than to theorem proving or standard conformance, since there are, in general, no algorithms for providing the solution to the problem to be solved. Thus, establishing an evaluation in such a setting is less direct.

For what concerns ontology matching evalutation most of the available works conveged towards contributing to the OAEI campaigns. Thus, below, we discuss the related

work on evaluation only in two relevant areas, namely semantic technologies in general and specifically database schema matching.

Evaluation of semantic technologies. While systematic evaluation of semantic technologies is not yet as established as in related areas, such as databases or information retrieval, several initiatives started to investigate this problem by focussing on different types of methods and tools. For example, early efforts have considered the evaluation of semantic web systems with respect to their ability of exchanging semantic data without loss of information [63]. Although in theory, interoperability should be granted by the use of standardized languages, such as RDF and OWL, evaluations have shown that this is not always the case. As a response to this problem, interoperability benchmarks for semantic web tools were defined and implemented for testing existing implementations [29]. So far, interoperability has mostly been tested for ontology development tools. More recent efforts also included the evaluation of APIs for ontology management and API-based interfaces [43].

The efficiency of accessing semantic data is another subject of existing evaluation efforts that stands in the tradition of database systems benchmarking, where the main focus has always been on efficiency. To this end, a number of benchmark datasets for evaluating the performance of RDF databases was defined in terms of generators that can be used to generate arbitrarily large RDF datasets based on a predefined schema [33; 6; 55]. The corresponding experiments typically focus on upload and query execution times. Compared to the existing benchmarking activities in the database area, a special characteristic of semantic data access is the need to perform logical reasoning for answering queries. This means that besides the efficiency, completeness and correctness of the underlying reasoning procedures are of a major importance and were also considered in the respective benchmarks, see e.g., [33; 44]. More recently, algorithms for generating test data that allows for measuring completeness of a reasoning system independent of a certain schema were investigated as well [61].

Another aspect of semantic technologies that was the subject of evaluation activities is the ability to find and combine relevant information in a useful way. Here, the main criterion is the quality of the resulting information. This task comes in different forms, depending on the kind of information that is concerned. While the use of semantic technologies for enhancing classical information retrieval tasks has not been the subject of systematic evaluation, there is some work from the area of web service discovery and composition, see, e.g., [66]. In particular, the task of selecting appropriate web services based on a user request and semantic annotations was investigated in detail and a comprehensive benchmarking suite is available [41]. Other benchmarking activities are concerned with the integration of different web services into a coherent workflow, although based on a qualitative evaluation rather than concrete quality measures [51].

Different communities have recognized the benefits of providing an automatic evaluation framework where system developers can test their tools against a predefined set of benchmark datasets and receive an evaluation result online. Examples are the SMT-Exec initiative[5] for satistisfiability testing and the S3 contest for web service matching[6]. The Ontology Alignment Evaluation Initiative described in this paper is a related

[5] http://www.smtexec.org

[6] http://www-ags.dfki.uni-sb.de/~klusch/s3/index.html

activity in the context of evaluating semantic technologies for finding and combining relevant information that focusses on the task of matching between knowledge models. It thus supplements, or has inspired, the activities mentioned above by focussing on a different technology.

Evaluation of schema matching. Untill recently there were no comparative evaluations and it was quite difficult to find two database schema matching systems evaluated on the same dataset. For example, an early evaluation effort of [16] focused mostly on comparison criteria from four areas, such as input (test cases), output (match results), quality measures (precision, recall, f-measure, overall) and savings of manual efforts (pre-match, post-match). It also provided a summary on several matching tools using those criteria. However, even at present in the database community there are no well-established benchmarks for comparing schema matching tools. Instead, the activities were somewhat fragmented, such as those of Cupid [45] and iMAP [15]. Several later works built up on the past results in terms of using the same datasets and quality measures for evaluations, such as COMA++ [3], S-Match [31], SMB [47] and YAM [19] to name a few. In turn, the work on STBenchmark [2; 1] focused on evaluation of mappings, namely on the transformation from source instances into target instances, what finds its parallels with the instance matching track of OAEI. The closest to OAEI works on benchmarking of database schema matching systems are those of [16] and more recently of XBenchMatch [18; 17]; though these initiatives have not led to well-established recurrent evaluation campaigns.

3 Ontology matching

Designing and running evaluation campaigns for a certain kind of tools require a solid understanding of the problem the respective tools try to solve. There have been different formalizations of the matching process and the results generated by this process [5; 42; 38; 59; 70]. We follow the framework presented in [27].

In order to illustrate the matching problem, let us consider two simple ontologies depicted in Figure 1. These ontologies contain subsumption statements, property specifications and instance descriptions. On an abstract level, ontology matching is the task of finding correspondences between ontologies. Correspondences express relationships supposed to hold between entities in ontologies, for instance, that a SubjectArea in one ontology is the same as a Topic in another one or that Regular author in an ontology is a subclass of Author in another one. In the example above, one of the correspondences expresses an equivalence, while the other one is a subsumption correspondence. In a further step, one may generate query expressions that automatically translate instances of these ontologies under an integrated ontology.

Matching is the process that determines an *alignment* A' for a pair of ontologies o and o'. There are some other parameters that can extend the definition of the matching process, namely: (i) the use of an input alignment A, which is to be completed by the process; (ii) the matching parameters, for instance, weights and thresholds; and (iii) external resources used by the matching process, for instance, common knowledge and domain specific thesauri.

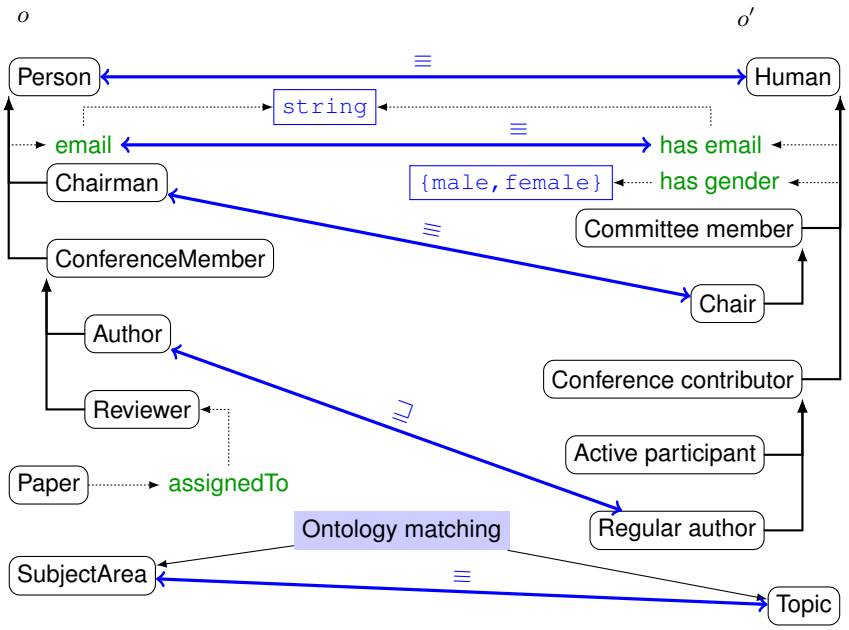

Fig. 1. Two simple ontologies. Classes are shown in rectangles with rounded corners, e.g., in o, Chairman being a specialization (subclass) of Person, while relations are shown without the latter, such as email being an attribute (defined on a domain string) and assignTo being a property. Ontology matching is a shared instance. Correspondences are shown as arrows that connect an entity from o with an entity from o'. They are annotated with the relation that is expressed by the correspondence.

Each of the elements featured in this definition can have specific characteristics which influence the difficulty of the matching task. It is thus necessary to know and control these characteristics (called dimensions because they define a space of possible tests). The purpose of the dimensions is the definition of the parameters and characteristics of the expected behavior in a benchmark experiment.

As depicted in Figure 2, the matching process receives as input three main parameters: the two ontologies to be matched (o and o') and, eventually, an input alignment (A). The input ontologies can be characterized by the input languages they are described (e.g., OWL-Lite, OWL-DL, OWL-Full), their size (number of concepts, properties and instances) and complexity, which indicates how deep is the hierarchy structured and how dense is the interconnection between the ontological entities. Other properties, such as consistency, correctness and completeness are also used for characterizing the input ontologies. The input alignment (A) is mainly characterized by its multiplicity (or cardinality, e.g., how many entities of one ontology can correspond to one entity of another one) and coverage in relation to the ontologies to be matched. In a simple scenario, which is the case for most of the OAEI test cases, the input alignment is empty. Regarding the parameters, some systems take advantage of external resources, such as WordNet, sets of morphological rules or previous alignments among general purpose resources, e.g., Yahoo and Google directories.

The output alignment A' is a set of correspondences between o and o':

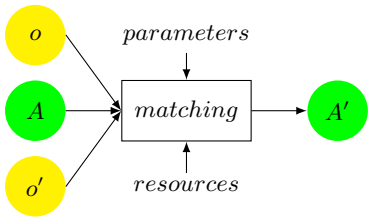

Fig. 2. The ontology matching process (from [27]).

Definition 1 (Correspondence). *Given two ontologies, o and o', a correspondence is a quintuple:*

$$\langle id, e, e', r, n \rangle,$$

such that:

- id *is an identifier of the given correspondence;*
- e *and* e' *are entities, e.g., classes and properties of the first and the second ontology, respectively;*
- r *is a relation, e.g., equivalence* ($\equiv$)*, more general* ($\sqsupseteq$)*, disjointness* ($\perp$)*, holding between* e *and* e'*;*
- n *is a confidence measure (typically in the* $[0, 1]$ *range) holding for the correspondence between* e *and* e'*.*

Alignments are sets of correspondences between entities belonging to the matched ontologies. The correspondence $\langle id, e, e', r, n \rangle$ asserts that the relation r holds between the ontology entities e and e' with confidence n. The higher the confidence, the higher the likelihood that the relation holds. For example, an alignment A, which contains only equivalence correspondences, is a 1:1 alignment, if for all $\langle id_1, e_1, e'_1, r_1, n_1 \rangle \in A$ there exists no $\langle id_2, e_2, e'_2, r_2, n_2 \rangle \in A$ with $(e_1 = e_2 \wedge e'_1 \neq e'_2) \vee (e_1 \neq e_2 \wedge e'_1 = e'_2)$.

For example, in Figure 1 according to some matching algorithm based on linguistic and structure analysis, the confidence measure between entities with labels Chairman in o and Chair in o' is 0.75. Suppose that this matching algorithm uses a threshold of 0.55 for determining the resulting alignment, i.e., the algorithm considers all pairs of entities with a confidence measure higher than 0.55 as correct correspondences. Thus, our hypothetical matching algorithm should return to the user the following correspondence $\langle id_{2,4}, \mathsf{Chairman}, \mathsf{Chair}, \sqsupseteq, 0.75 \rangle$.

Different approaches to the problem of ontology matching have emerged from the literature [27]. The main distinction among them is due to the type of knowledge encoded within each ontology, and the way it is utilized when identifying correspondences between features or structures within the ontologies. *Terminological* methods lexically compare strings (tokens or n-grams) used in naming entities (or in the labels and comments concerning entities), whereas *semantic* methods utilize model-theoretic semantics to determine whether or not a correspondence exists between two entities. Some approaches may consider the *internal* ontological structure, such as the range of the properties (attributes and relations), cardinality, and the transitivity and/or symmetry of

the properties, or alternatively the *external* ontological structure, such as the position of the two entities within the ontological hierarchy. The instances (or extensions) of classes could also be compared using *extension*-based approaches (e.g., based on frequency distributions). In addition, many ontology matching systems rely not on a single matching method (matcher), but combine several matchers.

4 Evaluation design

The design of the evaluations is at the heart of an evaluation campaign, and the design of a good evaluation is a task that should not be underestimated. Setting new challenges for participants in terms of well-designed tests requires a good understanding of the problem domain, in our case ontology matching. In fact the evaluation initiative only really took off after a theoretical framework for ontology alignment was developed within the KnowledgeWeb network of excellence [7]. Over the years, the theoretical understanding of the problem has been further improved and led to the development of further datasets.

Designing an evaluation is difficult, because it has to balance several partially conflicting desiderata:

D1: The evaluation criteria and tests should cover all relevant aspects of the problem and the results of an evaluation should provide a good estimation of the expected performance of the tested system in a real application.
D2: The evaluation has to be fair in the sense that it does not favor a certain approach or systems that make a certain assumption on the nature of the data or the result.
D3: The results have to be informative in the sense that they allow the developers of the tested system as well as potential users to learn about the strengths and the weaknesses of a tool and also to decide which tool shows a better performance.
D4: The evaluation should allow for quick feedback cycles to foster advances of the state of the art. This requires that the effort of conducting the campaign should not be too high neither for the participants nor for the organizers.

In the development of the Ontology Alignment Evaluation Initiative we have worked with these desiderata and came up with different methods for improving the evaluations to better meet them. These and further necessary developments are discussed in this section. We start with a basic evaluation design and then discuss its variations.

Figure 3 shows a basic evaluation process for ontology matching tools. The main component in this process is the *matching* component, which represents the system to be evaluated. The system takes two ontologies as input and generates an alignment. The second component is an evaluation script (*evaluator*) that takes the produced alignment and compares it with a reference alignment representing the expected outcome of the matching process. The evaluator compares the two alignments and computes a measure of the quality of the alignment produced by the matching component.

This basic process is simplistic and has to be concretized in many respects. First of all, the input data in terms of the ontologies to be matched has to be defined. No single pair of ontologies can test all aspects of ontology matching. We also experienced

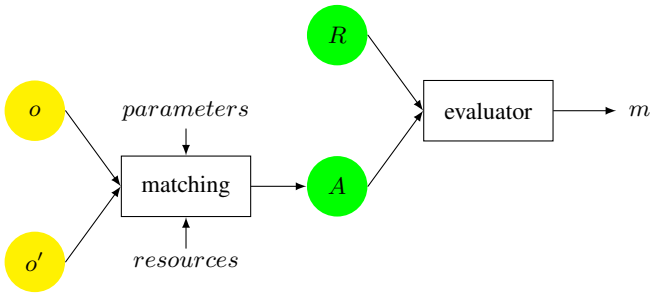

Fig. 3. Basic evaluation design: a matcher receives two ontologies o and o' as input and generates an alignment A using a certain set of resources and parameters. An evaluation component receives this alignment and computes a (set of) quality measure(s) m – typically precision and recall – by comparing it to the reference alignment R.

that there is a need for different types of datasets: for systematic evaluations and for competitive evaluations. Another insight gained was that standard quality measures, in particular precision and recall, are not always suited for the purpose of ontology matching as they fail to completely capture the semantics of ontology alignments and different measures are needed for evaluating different aspects. Finally, we found out that more complex approaches are sometimes needed in certain situations, for instance, if a partial alignment exists or if no reference alignment is available.

It is possible to use external resources as long as they have not been tuned to the current evaluation experiment (for instance, using a sub-lexicon, which is dedicated to the domain considered by the tests). It is acceptable that the algorithm prunes or adapts these resources to the actual ontologies as long as this is in the normal process of the algorithm. Moreover, some parameters can be provided to the methods participating in an evaluation. However, these parameters must be the same for all the tests. It can be the case that some methods are able to tune their parameters depending on the presented ontologies. In such a case, the tuning process is considered to be part of the method.

In the following, we elaborate these insights with respect to datasets, quality measures and evaluation processes used in the context of OAEI. Specifically, in §4.1, we discuss properties of ontologies and alignments that determine the hardness of a test. The datasets used in the OAEI initiative are presented in §4.2. In turn, §4.3 discusses evaluation measures and processes that were developed and used in OAEI. Finally, typical evaluation processes are discussed in §4.4.

4.1 Dataset characteristics

Good datasets are a prerequisite for a good evaluation. The nature of the datasets determines how far the evaluation design meets our first two desiderata: the coverage of relevant aspects and the fairness of the evaluation. In the case of ontology matching, a dataset typically consists of at least two ontologies and a reference alignment between these ontologies. In the following, we call the combination of exactly two ontologies and, if present, a reference alignment between these ontologies a *test*. A dataset consists

of several tests. If not defined otherwise, we assume that each combination of ontologies plus the respective reference alignment is a test in the dataset.

The work in [30] proposed the following criteria for designing or selecting datasets for ontology matching evaluation:

- Complexity, i.e., that the dataset is hard for state of the art matching systems.
- Discrimination ability, i.e., that the dataset can discriminate sufficiently among various matching approaches.
- Incrementality, i.e., that the dataset allows for incrementally discovering the weaknesses of the tested systems.
- Monotonicity, i.e., that the matching quality measures calculated on subsets of gradually increasing size converge to the values obtained on the whole dataset.
- Correctness, i.e., that a reference alignment is available for the dataset, which allows to divide generated correspondences into correct and incorrect ones.

There are two basic properties that determine the nature of a dataset, and thus, how well it meets the quality criteria mentioned above: the properties of the ontologies to be matched and the properties of the reference alignment, that are expected to be reproduced by the matching systems.

Ontologies. There are two major aspects of an ontology that have an influence on the matching process: the complexity of labels used to describe classes, relations and instances in the ontology, that has an influence on the initial determination of candidate correspondences, and the complexity of the structures used to define these elements that is often used to improve and validate the initial hypotheses.

Complexity of labels. Many matching systems use a combination of heuristics for comparing the labels of entities in ontologies in order to compute correspondences between these entities. Hence, the kind of labels found in an ontology influences heavily the performance of a particular matching system. Specifically, we distinguish between simple labels vs. sentence-like labels, monolingual vs. multilingual labels. It also often makes a large difference whether labels used in an ontology can be anchored to common background knowledge sources, such as WordNet, that helps interpreting those labels. Further complexity is added if the ontologies to be matched use specific vocabularies, e.g., from the biomedical or geo-spatial applications, that are outside common language.

Complexity of structures. Almost all matching systems use the structure of definitions in the ontologies to be matched in the later stages of the matching process to propagate similarity estimations and to validate hypotheses on correct correspondences. Therefore, structures found in ontologies are also an important issue in the design of benchmark datasets. Fortunately, the standardization of the semantic web languages RDF and OWL provide a common syntax for comparing ontologies, but still the way and intensity this common syntax is used varies a lot. Directories and thesauri only use the hierarchical structure given by subsumption, while more expressive ontologies use relations between classes that may be constrained by various kinds of axioms. This additional knowledge can be used by matchers for matching as well as for checking the coherence of their alignments [48].

On the level of instances, we can also have different levels of complexity. In particular, instances can either be described in detail using attributes and relations to other instances or can be atomic entities with no further explicit definitions or property specifications. Often instances represent links to external sources, e.g., web pages or images, that can be used as a basis for matching. In this case, the nature of the external resource can also make a significant difference. For example, web pages often provide a good basis for extracting additional information about the described object that makes matching easier, an image is harder to interpret and to compare with other resources.

Reference alignments. A reference alignment is another important aspect to consider: characteristics, such as the types of semantic relations used in the alignment or the coverage of the alignment, have a significant impact not only on the hardness of the task but also puts restrictions on evaluation measures that are discussed later.

Types of semantic relations. As mentioned in §3, an alignment consists of a set of correspondences defined by elements from the two ontologies and a semantic relation between them. The kind of semantic relations found in the reference alignment also determine what kind of relations the matching systems should be able to produce. The most commonly used relation is equivalence of elements (in most cases classes and relations). The majority of available matching systems are designed to generate equivalence statements. There are exceptions to this rule, however, that should be taken into account. Other kinds of relations that were investigated are subclass [67; 32] and disjointness relations [54; 32].

Formal properties of the alignment. Besides the type of a relation, its semantics is another relevant aspect. In particular, we have to distinguish between more and less rigorous interpretations of relations. The equivalence relation, for example, can be interpreted as logical equivalence or more informally as a high level of similarity or exchangeability. Using a rigorous formal interpretation of the semantic relations has the advantage that we can enforce formal properties on the reference alignment. For example, we can claim that the merged model consisting of the two ontologies and the alignment should be coherent, i.e., it should not contain unsatisfiable classes. Enforcing such consistency conditions is not possible for less formal interpretations.

Cardinality and coverage. A less obvious property with a significant influence on the evaluation results is the cardinality of the reference alignment. In principle, there is no restriction on the alignment, so the relation between elements from the different ontologies can be an n-to-m relation. In practice, however, it turns out that the alignment relation is one-to-one in most cases. Therefore, matching systems often generate one-to-one alignments. Along the same lines, the degree of overlap between the ontologies to be matched is not restricted and a dataset could consist of two ontologies with little or no overlap. Typically, however, it is assumed that the two ontologies to be matched describe the same domain. As a consequence, matching systems normally try to find a correspondence for every element in the two ontologies rather than ignoring elements.

4.2 OAEI datasets

From 2005 on, different datasets have been used in the OAEI evaluation campaigns. The aim of using these different sets is to cover as much as possible the relevant aspects of the matching problem, i.e., the desideratum D1 discussed above.

Initially, the goal of the initiative was to achieve this coverage within a single dataset, the *benchmark* dataset. The benchmark dataset deals with the topic of scientific publications. It consists of a large set of artificial tests. These tests alter an initial ontology and the task is to match it to the modified ontology. Modifications concern both the element labels, e.g., replacing them by random labels, and the structure, e.g., deleting or inserting classes in the hierarchy. In addition, the dataset comprises four other real ontologies that have to be matched to the reference ontology. Details about the different tests can be found on the OAEI website[7].

The declared goal of the benchmark dataset is the analysis of matching systems to identify their strengths and weaknesses with respect to the absence or the presence of certain structures in the ontologies to be matched. While the benchmark dataset serves this purpose quite well, it turned out to be less useful for other purposes. In particular, the benchmark dataset is not really suited for comparing the overall performance of systems. Obviously, comparing the performance of systems on the artificial tests is not useful for assessing system behavior in reality as each of the tests focuses on a specific situation that is not likely to occur in practice and the tests did not reflect any realistic situation. In consequence, we recognized that we needed other, more realistic tests to actually compare the performance of matching systems in realistic situations and that the benchmark dataset is not a suitable means for assessing matcher behavior on real tasks. However, it can be still used as an immediate first check-up of the newly proposed system in terms of its weaknesses, strengths and its presumable position with respect to the state of the art. Based on these experiences, the benchmark dataset was complemented by a number of other datasets that try to cover those aspects not addressed by the benchmark dataset. These datasets fall in different categories; see Table 1 for an overview of the datasets that are currently used in OAEI.

Expressive ontologies. For addressing the issues of realism and difficulty identified on the benchmark dataset, we have introduced two datasets that are more challenging in the sense that they are much larger, more heterogeneous and feature more complex definitions of classes that have to be taken into account during matching. The datasets in this category are the OntoFarm[8] dataset [69] also referred to as the *conference* dataset in the context of the OAEI campaigns and the *anatomy* dataset. The conference dataset consists of a set of fifteen OWL ontologies describing scientific conferences using complex definitions. The anatomy dataset consists of two ontologies describing the human and the mouse anatomy that are actually used in the medical community and have been manually matched by medical experts. For both datasets, reference alignments exist, but we have decided not to publish these reference alignments completely to avoid the effect we have observed for the benchmark dataset. Thus, it is possible to conduct a blind

[7] `http://oaei.ontologymatching.org/2009/benchmarks/`
[8] `http://nb.vse.cz/~svatek/ontofarm.html`

Dataset	Formalism	Relations	Confidence	Modalities	Language
benchmarks	OWL	=	[0 1]	open	EN
anatomy	OWL	=	[0 1]	blind	EN
conference	OWL-DL	=, <=	[0 1]	blind+open	EN
directory	OWL	=, <, >, $\perp$	1	blind+open	EN
library	SKOS +OWL	exact-,narrow-, broadMatch	1	blind	EN+NL+FR
benchmarksubs	OWL	=,<,>	[0 1]	open	EN
ars	RDF	=	[0 1]	open	EN
tap	RDF	=	[0 1]	open	EN
iimb	RDF	=	[0 1]	open	EN
vlcr	SKOS +OWL	exact-, closeMatch	[0 1]	blind expert	NL+EN

Table 1. Characteristics of test cases ('open' evaluation is made with already published reference alignments, 'blind' evaluation is made by organizers from reference alignments unknown to the participants and 'expert' evaluation involves manual analysis of results, by an expert user).

evaluation, where the correct answers are not given to the participants. Both datasets have become an integral part of the OAEI campaigns.

Directories and thesauri. These datasets consist of large weakly structured ontologies, as they are already in use on the web and in digital libraries. The lack of a sophisticated structure puts the element labels in a much more prominent position. Besides the analysis of labels, the size of the datasets in this category is a major challenge for many matching systems as the structures to be matched contain up to hundreds of thousands of classes. A problem connected to these more realistic datasets, e.g., *library* in Table 1, is lack of complete reference alignments. Due to the size of the models creating such an alignment manually is not an option, therefore other means of evaluation had to be found [36; 30].

Instance matching. With the increasing interest in linked open data, it turns out that typical matching problems on the web consist of finding instances representing the same individual rather than finding equivalent classes in different ontologies. While instance matching is covered by the theory of ontology matching outlined in §3, it has not been represented in the OAEI campaigns until recently. Since 2009 a number of instance matching datasets have been included in the campaigns. These datasets are the *iimb*, the *ars*, and *tap*. These datasets comprise automatically generated benchmarks, in which one dataset is modified according to various criteria, as well as real data from the domain of scientific publications.

Beyond equivalence. Finally, there are first attempts to move ahead from equivalence as the only semantic relation considered in the OAEI tests. There are tests now that ask for close matches as well using the relations 'exactMatch' and 'closeMatch'.

4.3 Evaluation measures

The diverse nature of OAEI datasets, mainly in terms of the complexity of test cases and presence/absence of (complete) reference alignments, has required to use different evaluation measures. Furthermore, evaluating a matching systems from different perspectives allows for avoiding to favor a certain approach or system, when evaluation is made under a same dataset. This is one of the criterion to meet the desideratum D2 presented above. Organizers have as well the important role of conducting a fair evaluation. Table 2 provides an overview of the evaluation criteria used in the OAEI evaluations.

Type	Compliance			Other			
Measure / Dataset	Manual labelling	Partial reference	Complete reference	Efficiency	Data mining	Logical Reasoning	Application oriented
benchmarks			√				
anatomy			√	√			
conference	√	√			√	√	
directory			√				
library		√					√
benchmarksubs			√				
ars			√	√			
tap			√	√			
iimb			√	√			
vlcr		√					

Table 2. OAEI evaluation criteria (compliance is usually measured with variations of precision and recall against available references).

The most commonly used and understood criterion for evaluation of ontology alignments is the compliance of matcher alignments with respect to the reference alignments. Measures, such as precision (true positive/retrieved), recall (true positive/expected) and f–measure (aggregation of precision and recall) have been used as basis in the OAEI campaigns for measuring compliance. For a subset of datasets, namely *conference*, *library* and *vlcr*, the complete reference alignment is not available and then compliance is measured on a partial reference alignment.

Although precision and recall are standard measures for evaluating compliance of alignments, alternative measures addressing some limitations of these measures have been used. For example, it may happen that an alignment is very close to the expected result (reference alignment) and another one is quite remote from it, although both share the same precision and recall. The reason for this is that standard metrics only compare two sets of correspondences without considering if these are close or remote to each other. In order to better discriminate such systems a relaxed precision and recall measures were defined which replace the set intersection by a distance [20]. To solve another problem, that two alignments may score differently while being semantically equivalent, semantic precision and recall were defined based on entailment instead of inclusion [22].

Specially in the cases where only partial reference is available, alternative evaluation approaches have been applied. For instance, in the conference track, manual labeling, data mining and logical reasoning techniques were considered:

- For manual labeling, for each matcher the most highly rated correspondences were considered as population. n correspondences per matcher were randomly sampled from the population. These correspondences were then evaluated as correct or incorrect. As a result, a score for precision was estimated.
- For supporting the discovery of non-trivial findings about matchers, data mining techniques and correspondence patterns were exploited as well. The aim is to find explanations on the so-called analytic questions, such as: (i) which systems give higher/lower validity than others to the correspondences that are deemed 'in/correct'?; (ii) which systems produce certain matching patterns/correspondence patterns more often than others?; and (iii) which systems are more successful on certain types of ontologies?
- Logical reasoning was used to measure the degree of incoherence that is caused by an alignment. The underlying idea is that a correct alignment should not result in unsatisfiable classes. Measuring the degree of (in)coherence of an alignment was first proposed in [48].

The approach adopted by the library track organizers, for compensating the lack of complete reference alignments, was based on application relevance. They considered the provided alignment in the context of an *annotation translation* process supporting the re-indexing of books indexed with one vocabulary A, using concepts from the aligned vocabulary B [36]. For each pair of vocabularies $A - B$, this scenario interprets the correspondences as rules to translate existing book annotations with A into equivalent annotations with B. Based on the quality of the results for those books for which the correct annotations are known, the quality of the initial correspondences can be assessed.

The criteria above are about alignment quality. However, another useful comparison between systems refers to their efficiency. The best way to measure efficiency is running all the systems under the same controlled evaluation environment. However, in the previous OAEI campaigns, participants were asked to run their systems on their own machines and to send the resulting alignments to be evaluated. So, the information about the time each system takes to execute the matching was gathered directly from the participants and could not be directly compared.

4.4 Evaluation processes

An evaluation process represents the interaction between several components in an evaluation experiment (matchers, test providers, evaluators, etc.). A simple process restricts the experiment to the evaluation of one matcher using a set of test cases.

Usually, several matchers are evaluated in one evaluation experiment. Figure 4 illustrates the evaluation process that extends the process presented at the beginning of this section (Figure 3). The first step is to retrieve, from a database of *tests* containing the ontologies to be matched and the corresponding reference alignments, the tests to

be considered in such an evaluation. Next, each available *matcher* performs the *matching* task, taking as input parameters the two ontologies. Then, the resulting alignment is evaluated against the reference alignment, by an *evaluator*. Finally, each result interpretation is stored into the *result* database (for instance, precision and recall).

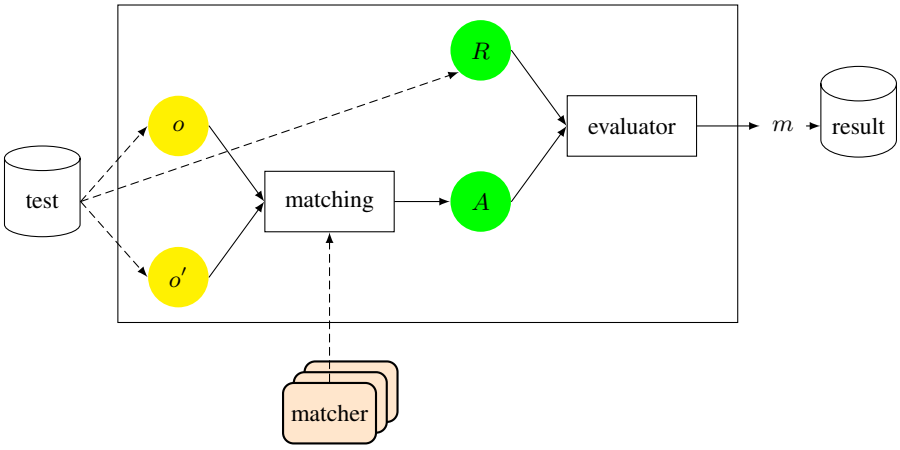

Fig. 4. Basic evaluation process.

Due to the variability of the alignment evaluation, different scenarios can be specified, by adding new components to the process presented in Figure 4:

Test generator. Test cases can be generated from a description of the kind of evaluation to be executed (for example, removing n% of the properties of the ontologies). A description of the desired test case must be provided and the output of the test generator service is then used as input to the matching process.

Lack of reference alignment. It is not the case that all test cases have a complete reference alignment. Thus, alternative evaluation metrics must be provided, such as measuring the consensus between several matchers, intersection or union of the results, etc.

User in the loop. Sometimes, matching systems are considered as semi-automatic and the user has control over the matching process. On the other hand, manual labeling can be required in the cases where the reference alignments are not available.

Task-specific evaluation. It can be useful to set up experiments which do not stop at the delivery of alignments, but carry on with the particular task. This is especially true when there is a clear measure of the success of the overall task; see §4.3.

The components described above can be combined together in different ways. Figure 5 illustrates a more elaborated process where tests are generated by a test generator, according to the description provided by the user. This generation process may create a set of alternative ontologies, from a reference ontology, by removing its properties or

individuals. Moreover, one can imagine that no reference alignments are provided by the test generator. In such a scenario, the user has the role of an evaluator. For each generated test, the available matchers are executed and their resulting alignments are stored into a database, whose content will be used later for user evaluation.

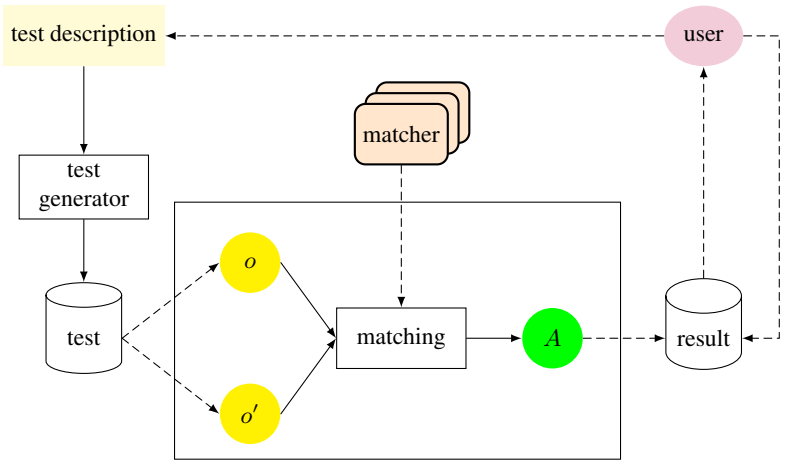

Fig. 5. Advanced evaluation process.

5 Analysis of the past OAEIs

The evaluation design presented in the previous section was chosen to provide tool developers and potential users with feedback on the state of the art in ontology matching and to foster developments in the field, meeting the two last desiderata presented in §4. Therefore, a crucial question that needs to be answered is whether the initiative indeed supported an improvement of the field. In the following, we try to answer this question by providing an abstract view on the results of the evaluation campaigns. This overview shows that OAEI was a success in many respects. First of all, a large and vivid community was established around the OAEI campaigns, which is shown by an increasing number of participants and test cases provided by the community. Further, we will show that there actually has been an improvement of matching systems that frequently participated in the benchmarking campaigns both in terms of runtime and quality of matching results. Finally, and probably most importantly, we have gained insights in the nature of the ontology matching tasks and the functioning of matching systems.

We first provide an overview of the evaluation campaigns that have been carried out from 2004 to 2010 (§5.1). We then summarize our observations with respect to the evolution of result quality (§5.2), the efficiency of matching systems (§5.3) and with respect to the impact of matching system configurations on the results (§5.4).

5.1 Campaigns and participation

The first ontology matching evaluations were carried out in 2004 as part of the Information Interpretation and Integration Conference (I3CON)[9] held at the NIST Performance Metrics for Intelligent Systems (PerMIS) workshop and the Evaluation of Ontology-based Tools (EON) workshop of the annual International Semantic Web Conference (ISWC) [62]. The workshops were organized in a joint but complementary way by different organizers. This parallel development emphasized the importance of the topic and indicated that a joint initiative would be of advantage. From 2005 on, joint activities are carried out under the heading of the Ontology Alignment Evaluation Initiative. The first official evaluation campaign of the initiative was carried out at the Workshop on Integrating Ontologies at the 3rd International Conference on Knowledge Capture (K-CAP 2005) in Banff, Canada. Since 2006, the annual evaluation campaigns are carried out at the International Semantic Web Conference in the context of the Ontology Matching workshop. Since the early beginning the workshop has a constant attendance of more than 50 participants working on the topic. Over the years, the number of systems participating has increased from 4 systems in 2004 to 15 in 2010. A detailed look shows that there was a significant increase from 4 in 2004 up to 17 in 2007, while from 2007 to 2010 the participation rate is with ≈ 15 participants relatively stable and fluctuates around this point. In future it is required to extend OAEI with new datasets and evaluation modalities according to the trends in the field (see §6) in order to maintain or increase the participation rate.

Table 3 provides an overview of the campaigns carried out so far. More information on the individual campaigns can be found on the OAEI web site[4].

	year	location	#tests	#participants	reference
I3CON	2004	Gaithersburg, US	10	5	-[9]
OAC	2004	Hiroshima, JP	1	4	[62]
OAEI	2005	Banff, CA	3	7	[28]
OAEI	2006	Athens, US	6	10	[26]
OAEI	2007	Busan, KR	7	17	[25]
OAEI	2008	Karlsruhe, DE	8	13	[8]
OAEI	2009	Chantilly, US	9	16	[23]
OAEI	2010	Shanghai, CN	6	15	[24]

Table 3. Overview of the evaluation campaigns.

5.2 Quality improvement

The main goal of OAEI is to support the enhancement of the ontology matching field. In the following, we report on results that show in how far this goal has been reached. First, we present summative results. In particular, we show how the average f-measure developed from 2005 to 2010 analyzing those datasets which have been run several

[9] `http://www.atl.external.lmco.com/projects/ontology/i3con.html`

years in succession. Then we analyze those systems that have been participating continuously from 2007 to 2010 in detail. The presented results allow to discuss the effects of a continuous participation.

Summative results. Our analysis required to recompute some of the values presented as results of the annual campaigns. In particular, the benchmark data was rendered more difficult in 2008. Since these changes affected both ontologies and resulting reference alignments, we did not recompute these values. The reference alignments of the conference track have been extended year by year. We recomputed the average f-measure based on the current, most comprehensive corpus of reference alignments. This has to be taken into account when analyzing the results. We have compared the average f-measure in terms of the arithmetic mean over all participants per year and track. This gives a rough representation on the main tendency and allows for abstracting from interdependencies between precision and recall.

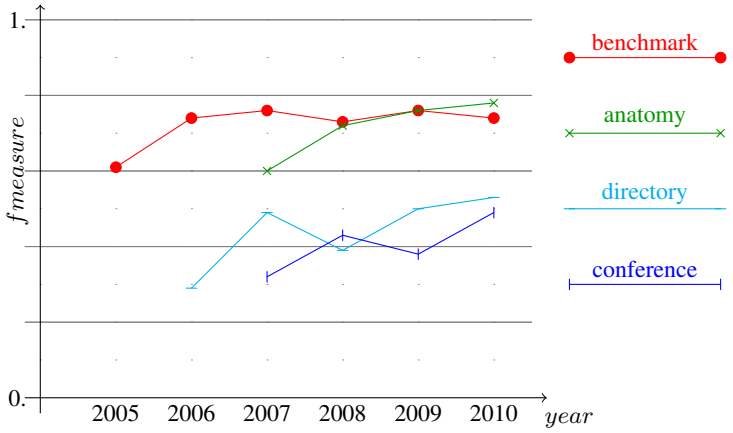

Fig. 6. Evolution of the average f-measure in different datasets.

The results of Figure 6 are heterogeneous for the different datasets. The results for the conference and directory dataset range from an f-measure of 0.3 to 0.5. Both datasets leave room for further improvements. A detailed look reveals that there is a high variance between participants. The top performers of the last years reached an f-measure in the range of 0.5 to 0.6 in the conference track and an f-measure of ≈ 0.6 in the directory track. The average f-measure for the benchmark and anatomy datasets ranges from 0.6 to 0.8, even though both datasets describe different domains and vary in size and expressivity. In both cases good results in f-measure are based on a high precision. The challenge with regard to these datasets is to increase recall of the results without decreasing the high precision scores.

We observe a moderate increase in benchmark and conference with some exceptions. Results for anatomy and benchmark will be analyzed in more detail later on. The quality of the alignments submitted to the conference track increases with each year, with the exception of 2008. However, Figure 6 does not show that the average f-measure

is based on very different results generated by each of the participating matching systems. It seems to be hard to find an appropriate configuration for matching ontologies of the conference dataset that is also appropriate for the other datasets. We discuss this issue in detail in §5.4. The improvements of the anatomy results are the most significant. In particular, we measured for each year a continous improvement. Remember that the reference alignment of the anatomy track was not available for participants (blind modality) until 2010 and it is hardly reconstructable without biomedical expertise. For what concerns the directory track (where the reference alignments were partially available), the overall trend from 2006 to 2010 is positive, though with a substantial drop in 2008. There are several explanations for this: (i) OLA2 and Prior+ never participated again after 2007 and those were the two systems showed top results, (ii) the set of participating systems in 2008 was almost completely different compared to 2007; it performed worse than the set of participating systems in 2007, but better than those participating in 2006. Overall we conclude that the field as a whole improved over the years.

We have also claimed that systems entering the campaign for several times tend to improve over years. By providing matching system developers with measurable feedback on their developments, it seems reasonable to think that they will be able to analyze the reasons for these results in order to improve their systems. We consider this claim in the following.

Very few of these systems have participated in most of the tests and also only a few systems have participated more than three years in a row, thus allowing a judgement of their individual improvement over time. We therefore have to base our discussion on quality improvement on a limited set of systems and datasets. From among the datasets that were systematically evaluated against a reference alignment from 2007 to 2010 we have chosen the benchmark and the anatomy datasets. We have selected these tracks because several systems participated at least three of four times in these tracks from 2007 to 2010. For the other tracks, we found a lower number of (more or less) continuous participation.

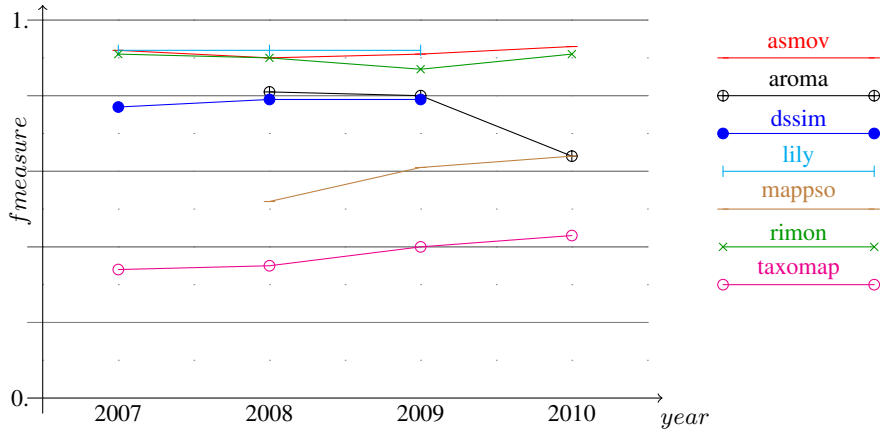

Fig. 7. Evolution of results on the benchmark dataset.

Results on the benchmark dataset. Figure 7 shows f-measure of the systems under consideration on the benchmark dataset in 2007 through 2010. These systems achieve a similar level of precision, between 80% and 95%, which is quite a high value for a matching task. Only recall differs and impacts f-measure. However, for each system there is little variation, and not necessary towards an increase. This is confirmed by the results of ASMOV and RiMOM which have participated for four years in a row, respectively .92 and .9 in f-measure.

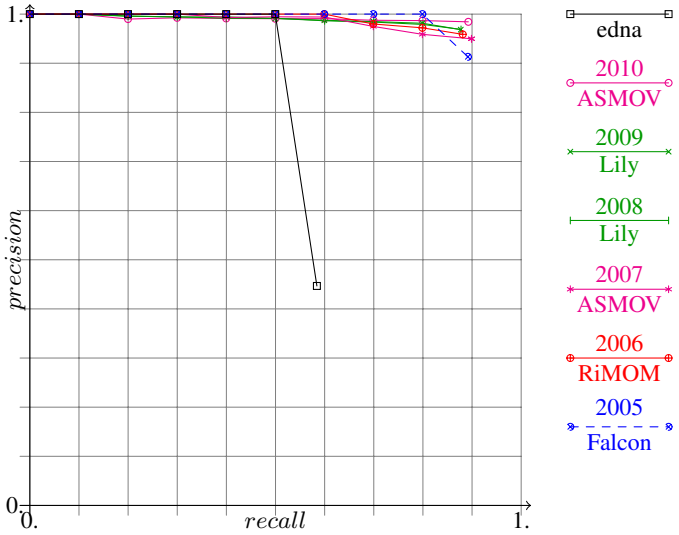

Fig. 8. Precision/recall curves for the yearly best results on the benchmark dataset (edna is a simple matcher based on edit distances on names of entities).

Figure 8 shows that the best systems are overall very safe because their precision is very close to 100% until 50% of recall and it is still over 95% until 90% of recall. They are also able to stop providing correspondences when their quality goes down (compared to the edna baseline). Figure 8 also shows the yearly progress of these systems by preserving better precision when looking for more recall.

The reasons for this behavior is that benchmark is made of a myriad of tasks, some of which are very difficult, but most of which are relatively easy. In consequence, the overall results are higher than for other tasks which were designed to be realistic or hard. This means that the results (precision, recall, f-measure) cannot be transposed from benchmarks to other situations. This also explains why gaining the last points of precision and recall is difficult for each system individually. Due to this feature, the benchmark dataset has lost its discrimination power over the years: a large difference between systems on the benchmarks still reflects a difference in the versatility of systems in practice, but small differences are not relevant. In addition, benchmarks are often used by system developers to tune their systems both because they cover a variety of situations and because reference alignments are available. Hence, even systems which participate for the first time achieve relatively high performances. This may lead systems to be overfitted to the benchmark dataset, i.e., tuned to solve this particular kind

of tasks. Only systems which are especially designed for another purpose and whose designers do not want to twist to address benchmarks achieve low performances.

In summary, although artificially designed, benchmarks can be used as a starting point for developers to test which kinds of ontology features their systems handle better. This feedback can be then exploited for further improvements in their implementations. However, it is not relevant for predicting the behavior of a system in a particular situation.

Results on the anatomy dataset. If our explanation of the results on the benchmark dataset is correct, and there is still an improvement of the overall quality of individual matchers, this improvement will have to be visible in the results on the blind datasets. We chose the anatomy dataset as a basis for checking this as it has been evaluated against a complete reference alignment from 2007 on. Further, we present the results of those systems that particpated at least three times in these four years.

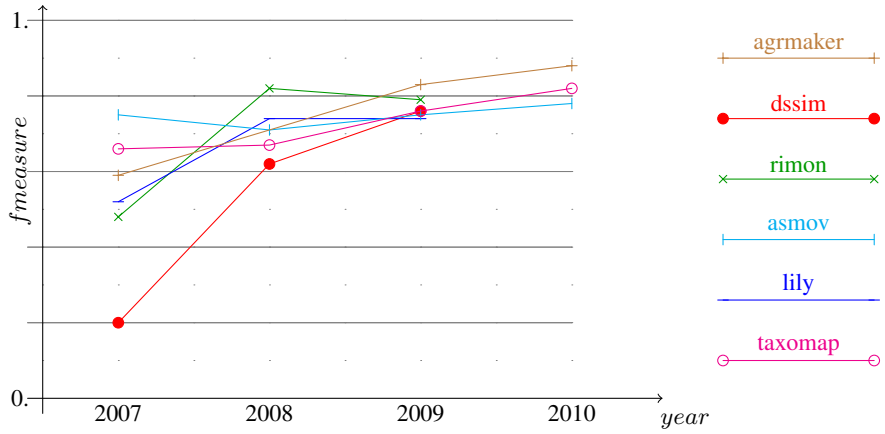

Fig. 9. Evolution of results on the anatomy dataset.

Figure 9 shows the development of f-measure of the systems from 2007 to 2010. This time, we can clearly see an upward trend in f-measure, which reflects both a significant increase in precision and a moderate increase in recall. This trend is more significant for the second time a system participates, reinforcing the analysis above that once participants know the results of the evaluation they can better improve it, but the next time the increase will be smaller. This pleads for more tests and more reference alignments given to participants because this can be a sign of over fitting to the OAEI datasets in general.

Hence, we conclude that there has been a significant increase in the quality at least of those matching systems on real world datasets that participated in the evaluation on a regular basis. This supports our claim that OAEI observes a measurable quality improvement in the ontology matching field.

5.3 Runtime

Besides the quality of generated alignments, other criteria are important for practical applications. With the increase of the number and the size of existing ontologies, the runtime of systems becomes also crucial. We therefore made first attempts of measuring the runtime of matching systems as well. This was done in a systematic way for the first time in 2007 evaluation of the anatomy track. Runtime was not a topic of investigations for the other OAEI tracks, thus we have to focus on the results for the anatomy track.

Due to the setting of previous OAEI campaigns, where system developers run their matchers on the test sets locally and send the results for inspection, it was not possible to measure comparable runtimes on a fair ground. For that purpose, it would have been necessary to execute all systems on the same machine ensuring a fair comparison. As an alternative, we collected statements about runtimes that have been measured by the participants themselves. This information had to be delivered together with a description of CPU and memory capabilities of the computer on which the matching process was executed. According to these descriptions, in 2009 most of the systems were run on a CPU with two cores in the range from 2.0 to 3.1 GHz, using 2 or 4 GB RAM memory.

In 2007, this survey was conducted by OAEI organizers mainly for interest. However, the huge variability in the reported runtimes together with the fact that all systems were executed on machines of similar strength, encouraged us to publish runtimes as part of the results in 2007, following the same strategy in 2008 and 2009. In 2010 it was originally planned to conduct the track in a completely automized evaluation setting. Finally, the evaluation was conducted in semi-automized way. Due to this, runtime measurements are unfortunately missing for 2010. Table 4 gives an overview on the runtimes that were reported by the matching tool developers.

System / Year	Anchorflood	AgreementMaker	Aroma	ASMOV	DSSim	Falcon-AO	kosimap	Lily	Prior+	RiMOM	SAMBO	SOBOM	TaxoMap	X-SOM	**Average**	**Median**
2007	-	30	-	900	75	12	-	5760	23	240	360	-	300	600	830	270
2008	1	-	4	230	17	-	-	200	-	24	720	-	25	-	152.6	24.5
2009	0.25	23	1	5	12	-	5	99	-	10	-	19	12	-	18.6	11

Table 4. Runtimes in minutes reported by the participants.

These results show that the community has made clear improvements regarding runtime issues. In 2007, one of the systems required four days for matching the ontologies, while the slowest system in 2009 finished the matching process in under two hours. This trend is also reflected by the average and median values that significantly decreased from 2007 to 2009. It is also interesting to see that the median and average seem to converge in 2009 because there is no longer negative outliers that require an enormous amount of time. Note also that the decreased runtime is in most cases not related to a decreased quality of the generated alignment.

More important than the trend line of faster matching systems, is the positive acceptance of presenting and discussing runtimes as part of the evaluation. Although we are aware that this way of gathering reported runtimes is a subject to criticism, the approach has nevertheless pointed to an important aspect in evaluating matching systems that would have been neglected otherwise. In §6.3, we describe an infrastructure that will finally allow to measure runtimes by the use of an evaluation runtime environment.

5.4 System configuration and optimisation

We study here the configuration of participating systems and the possible influence of datasets and evaluation settings on the performance of systems. For that purpose, we examine the blind results obtained on the conference dataset with the results obtained by the same systems in the benchmark dataset.

Our analysis is based on a discussion of the results from the OAEI 2009 conference track. All of the participants of the conference track have also participated in the benchmark track, most of them with good results. These systems are AFlood [56], AgreementMaker [11] as well as an extension of the system, AROMA [13], Asmov [37], Kosimap [52] and DSSim [50]. Two systems, namely DSSim and AFlood, did not annotate correspondences with confidence values. Since our approach requires confidence values, we omitted these systems.

The conference submissions of 2009 are well suited for our analysis, because in 2009 the evaluation of the submitted results was based for the first time on the use of a substantial set of reference alignments. Only a small subset of these alignments has been available prior to the evaluation. Similar to the ontologies of the benchmark track, the ontologies of the conference track are of moderate size (between 32 and 140 classes). They cover the domain of conference organization. This domain is partially overlapping with the domain of bibliography. In addition, in both cases, ontologies are labeled in natural language terms (as opposed to the ontologies of the anatomy track, for instance).

Thus, we would expect that a system that obtains good results for the benchmark track, obtains similar results for the conference track. In particular, we would expect that the configuration of such a system is also well suited for the conference track. However, the results do not fit with this hypothesis.

In Figure 10, the dependency between f-measure and threshold is shown for each system that participated in this track. Figure 10 is generated, for each submitted alignment featuring confidences different than 1, by applying a posteriori a threshold that is increased step by step. For each setting the resulting f-measure is measured and its average for all test cases is depicted in Figure 10.

For each of the other systems, we can distinguish between two interesting points. The first one is the threshold t where the f-measure increases for the first time (e.g., for kosimap $t = 0.1$, ASMOV did not use a threshold). This point refers to the threshold that was chosen by the tool developer when running his matching tool to generate the alignments. Since none of the correspondences has a confidence value below this threshold, we observe a horizontal line for, e.g., $t < 0.1$ with regard to kosimap. The second interesting point is the threshold t where the system reaches its maximum f-measure (e.g., for kosimap $t' = 0.52$).

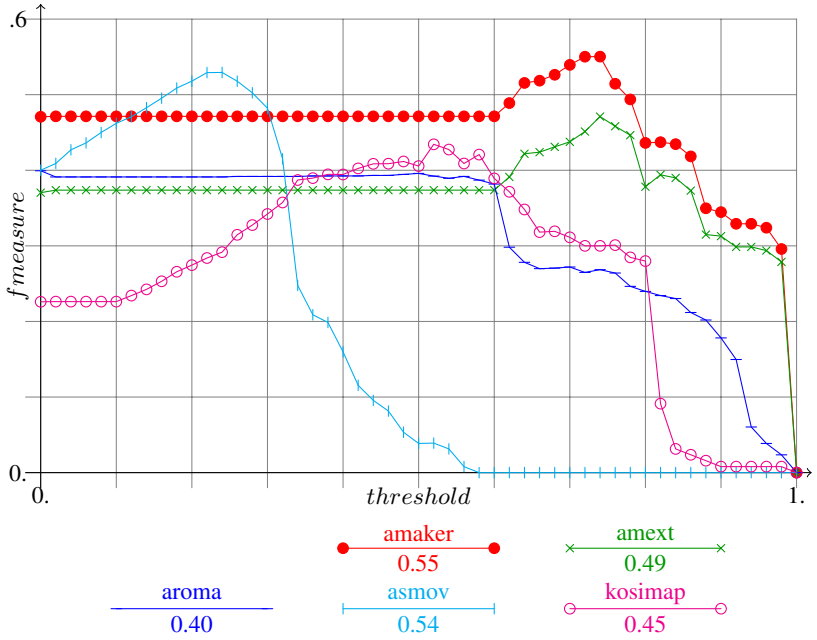

Fig. 10. F-measures of matching systems for different thresholds (the values under the legends are the absolute optimal f-measure for each system).

These curves are all nearly monotonically increasing until the optimum and then monotonically decreasing. This could be the sign of a robust way from all systems to rank correspondences (the increasing phase corresponds to less than optimally ranked false positives and the decreasing phase, more than optimally ranked true positives). On the other side, if these systems are good at ranking correspondences, they are not very good at finding the optimal threshold. Moreover, they are all lower than optimal.

However, the f-measure of these systems is far lower than the one they obtain in the benchmark track. How can this be explained? We cannot exclude that systems are overfitting on benchmarks, i.e., they are optimized for performing well at benchmark. Alternatively, the benchmark dataset has a particular feature that can favor those systems which try to maximize pairing of entities, e.g., by considering a similarity between two entities as a gain and to maximize the gain. Given one of the benchmark test cases, let o_s be the smaller ontology and let o_l be the larger ontologies. Then each matchable entity in o_s has a counterpart in o_l. Thus maximizing pairing is a good strategy. We call this over-matching. In fact, overfitting or over-matching would have the same results.

This characteristic occurs in most of the benchmark tests and in none of the conference tests. Moreover, it is only likely to occur in specific scenarios such as version matching. So, it introduces a bias in evaluation results towards a particular strategy.

The lessons of this analysis from the standpoint of evaluation are three fold:

- The benchmark test case should be modified so that this bias is suppressed;
- Delivering more datasets with reference alignments would help developers avoiding overfitting;

– Multiplying datasets and studying divergence is a good direction because it allows to test matchers in different situations and cross-compare the results from multiple datasets.

Further analysis that goes beyond the scope of a single track is required to understand the effects and appropriateness of specific system configurations.

6 Trends and challenges

From the last years of OAEI (2005-2010), we can identify medium term trends and challenges for ontology matching evaluation. We distinguish between two types of trends: some trends are relevant for the evaluation of systems in general (§6.1) while others are more specific to the matching problem (§6.2). The first ones are independent of the concrete problem that is solved by a system to be evaluated. Typical examples are evaluation metrics related to the effectiveness of user interaction or issues related to the hardness of a test. In addition to these two groups of trends, we finally conclude in §6.3 with a challenge that is tangent to many of the issues. We discuss the automation of ontology evaluation and the infrastructure required for that purpose. In particular, we present the infrastructure developed in the SEALS project[10] (Semantic Evaluation at Large Scale), which represents a partial continuation of OAEIs, as key to solve many open issues.

The work in [60] described ten challenges for ontology matching. Amongst others, these include *large-scale evaluation*, *performance of ontology matching techniques* and *reasoning with alignments*, which are directly related to trends identified here.

6.1 General issues

User interaction. Automatic ontology matching can only be a first step in generating a final alignment. Therefore, systems that (i) automatically generate an initial set of matching hypothesis, (ii) support the user in the refinement of the generated alignment, (iii) propagate the user input to semi-automatic filter and/or extend the alignment are in advance and will finally be used to solve concrete matching problems. An example for such a system is AgreementMaker [10]. However, current evaluation techniques do not take into account the quality and effectiveness of user interventions. Currently, only a subtask of the anatomy track in OAEI deals with the third aspect marginally, while the second point is not at all considered. This is one of the most important drawbacks of current evaluation practices that has to be tackled in future.

In situ evaluation. The overall approach underlying most OAEI evaluations is based on the implicit assumption that there exists a unique reference alignment that correctly describes how ontologies have to be matched. Although this reference alignment is not always available, correspondences can in principle be divided in correct and incorrect ones. However, the relative quality or usefulness of a generated alignment also depends on its intended use. The difference between these approaches was emphasized in [68]

[10] `http://about.seals-project.eu/`

by a comparison of *relevance* and *correctness*. In [34], an evaluation method is described that takes into account some characteristics of a usage scenario and reports the respective evaluation results.

Large scale analysis. OAEI campaigns gave only some preliminary evidence of the scalability characteristics of the ontology matching technology. We reported about these attempts in §5. Therefore, larger tests involving 10.000, 100.000 and 1.000.000 entities per ontology (e.g., UMLS has about 200.000 entities) are to be designed and conducted. In turn, this raises the issues of a wider automation for acquisition of reference alignments, e.g., by minimizing the human effort while increasing an evaluation dataset size [60; 46]. Notice also that scalability involves not only the consideration of runtime, but has to focus also on aspects as memory consumption and required disk capacity.

Defining and measuring test hardness. There is a need for evaluation methods grounded on a deep analysis of the matching problem space. Semi-automatic test generation methods require such an analysis as basis. These methods will allow for the construction of tests of desired hardness by addressing a particular point in the matching problem space. We have already argued that additional tests are required. Initial steps towards this line were already discussed in [30].

6.2 Specific issues

In the following, we present several specific issues that we believe will become more important to OAEI: complex matching, instance matching and database schema matching. Complex matching refers to a matching process in which correspondences are not restricted to link named entities, but can also link complex descriptions. Instance matching is not concerned with matching terminological entities but focuses on matching individuals. Finally, schema matching has received decades of attention in the database community. Database schemas are different from ontologies, e.g., by not providing explicit semantics for their data. However, these are also similar in the sense that both schemas and ontologies provide a vocabulary of terms and constrain the meaning of terms used in the vocabulary. Moreover, in real life situations schemas and ontologies have both well defined and obscure labels and structures, thus, these often share similar solutions, which need to be evaluated.

Complex matching. State of the art ontology matching techniques are often limited to detect correspondences between atomic concepts and properties. Nevertheless, for many concepts and properties atomic counterparts will not exist, while it is possible to construct equivalent complex concept and property descriptions [58]. A typical example, presented in [53], is the correspondence $\mathit{Researcher} \equiv \mathit{Person} \sqcap \exists \mathit{researchedBy}^{-1}.\top$. The expressivity supported by the available Alignment API [21] implementation was in the past restricted to non-complex correspondences and has recently been extended to a more expressive language referred to as *EDOAL* (Expressive and Declarative Ontology Alignment Language) [14]. Even though the infrastructure for expressing complex correspondences is now available and several approaches for complex matching techniques have been proposed (see, for example, [15; 53]).

Instance matching and linked data. While rich ontologies were promoted as an integral part of every semantic web application [35], it is increasingly argued that the real value of the semantic web is based on its ability to create and maintain linked open data which provides effective access to semantically enhanced information on the web [65]. In 2009, OAEI comprised for the first time a track explicitly concerned with instance matching. In 2009 six matching systems participated, in 2010 five systems participated. It can be expected that this track will be an important component of the OAEI in the following years with an increasing number of participants.

Database schema matching. As was mentioned in §2, at present in the database community there are no well-established benchmarks for comparing schema matching tools. However, there are many recent schema matching tools and more generally model management infrastructures, e.g., COMA++ [3], AgreementMaker [12], GeRoMe [40; 39], Harmony [49; 57], that are able also to process ontologies, and hence, might be interested to test them within OAEI, as actually already happens, though modestly. On the other hand, OAEI has to consider including explicit schema matching tasks involving XML and relational schemas in order boost the cross-fertilization between these communities.

6.3 Automation

Although OAEI campaigns have created a basis for evaluation that did not exist before, the progress in leveraging increased evaluation efforts has to be made in order to continue the growth of ontology matching technology. Further progress is highly dependent on the automation of many parts of the evaluation process. This would reduce the effort necessary for carrying out evaluation, but above all, this would allow to handle more complex evaluation processes as well as measurements of runtime and memory consumption. Reducing the evaluation effort will allow for better meeting the fourth desideratum discussed in §4.

The SEALS project aims at establishing systematic evaluation methods for semantic technologies, including ontology matching, by providing standardized datasets, evaluation campaigns for typical semantic web tools and, in particular, a software infrastructure – the SEALS platform – for automatically executing evaluations. This platform will allow matcher developers to run their tools on the execution environment of the platform in both the context of an evaluation campaign and on their own for a formative evaluation of the current version of the tool. The results can be published both in the context of the evaluation campaign or in the context of evaluating a tool on its own. In both cases, results are reproducible, since the matching system, the test dataset and the results themselves are stored and archived in the repositories of the SEALS platform.

This approach differs from the approach conducted in the OAEI campaigns where participants send their results (and their systems) to the OAEI organizers in the Alignment API format [21]. These submissions are accepted by the organizers as official results of the matching system. After a phase of validating, e.g., the format of the submissions, evaluation experiments are conducted by the organizers and the results are

prepared and finally presented on a webpage[11] and in the annual result reports [28; 26; 25; 8; 23; 24]. This process requires several weeks before first results are published.

The SEALS platform aims at automating most of the evaluation process. This allows tool developer to receive a direct feedback. OAEI will in particular benefit from both the reduced amount of effort required by the organizers and from the controlled execution environment. This environment ensures that the matching systems generate the alignments with a fixed setting for each track and test case. In particular, it allows to execute all evaluated matching systems in the same controllable context. Thus, it is possible to conduct precise runtime measurements that will replace the report-based approach used from 2007 to 2009.

OAEI and SEALS are closely coordinated and the SEALS platform will be progressively integrated within the OAEI evaluations. In a first phase, the participants of three OAEI 2010 tracks (benchmarks, anatomy, conference) were asked to make their tools available as web services. Implementing the required interface allowed participants in 2010 to debug their system from their own site. This approach substitutes the phase of preliminary testing as well as the submission of the final results. The alignments are generated on the machine of the tool developer and sent to the SEALS platform in the context of an evaluation process. On the one hand, evaluation results are immediately available in this setting. On the other hand, runtime and memory consumption cannot be correctly measured due the fact that the controlled execution environment is missing. Details on this approach can be found in [64].

In the second phase, which is planned already for OAEI 2011, the tools will be deployed in the SEALS platform. This allows organisers to compare systems on the same basis, in particular in terms of runtime. This is also a test of the deployability of tools. The successful deployment relies on the Alignment API and requires additional information about how the tool can be executed in the platform and its dependencies in terms of resources (e.g., installed databases or resources like WordNet). For that reason, the challenging goal of the SEALS project can only be reached with the support of the matching community and it highly depends on the acceptance by the tool developers.

7 Conclusions

The OAEI campaigns of the last years have provided extensive experience in ontology matching and evaluation of semantic technologies in general. This experience was reported in this paper. We summarize lessons learned that are worth emphasizing because they are relevant not only to OAEI, but also to the evaluation activities in other areas of semantic technologies and beyond.

As the reported experience indicates, foremost, there is a real need for systematic evaluation. Researchers and practitioners of the ontology matching tools have eagerly taken up the challenges offered by OAEI and actively participated from the beginning on. In general, systems have improved their performances over the campaigns for most of the tracks. This is specially corroborated by the results for the anatomy track, but this is a general trend.

[11] See, for example, `http://oaei.ontologymatching.org/2010/results/`

We observed that it was necessary to evolve with the field, involving our understanding of it and the reaction of developers to the proposed datasets. For example, most of the participants focused on the benchmark dataset, followed by anatomy and conference. There are only few systems that did not submit their results for benchmark. It can be due to the fact that benchmark offers relatively easy tasks and full reference alignments. Developers naturally use this available information (evaluation results) for improving their results. However, this overfitting has a potential influence on the performance of the systems. In turn, this requires to be reactive in proposing new datasets, new measures and new evaluation settings. We have pointed out areas in which improvements are necessary: more varied benchmarks (from various vertical domains as well as transversal ones), instance-based and user-assisted matching to name a few.

Also we made the case for automation and reported about first steps made in that direction. Increased automation does not only mean less work for evaluation organizers and better reproducibility. It offers the opportunity to generate datasets and thus to test scalability, variability and to understand test hardness. This allows for performing runtime, space and deployability measurements. Ultimately, it turned out that a rather minimal common infrastructure was sufficient to start the initiative.

Finally, setting up such an evaluation is a great chance, and a great responsibility: it has an influence not only on the improvement of systems but also on research directions being followed. This chance, however, comes at a price, since the successful evaluation initiative requires a deep understanding of the problem domain and substantial resources to be dedicated to creating datasets, designing protocols and processing evaluations.

Acknowledgements

Jérôme Euzenat, Christian Meilicke, Heiner Stuckenschmidt and Cássia Trojahn dos Santos are partially supported by the European project SEALS (IST-2009-238975). Pavel Shvaiko was supported by the Trentino as a Lab initiative of the European Network of the Living Labs at Informatica Trentina.

We are grateful to all our colleagues who contributed to the OAEI campaigns: Caterina Caraciolo, Alfio Ferrara, Ryutaro Ichise, Antoine Isaac, Fausto Giunchiglia, Willem Robert van Hage, Laura Hollink, Cliff Joslyn, Véronique Malaisé, Malgorzata Mochol, Andriy Nikolov, Natasha Noy, Juan Pane, Marta Sabou, François Scharffe, Vassilis Spiliopoulos, Ondřej Šváb-Zamazal, Vojtech Svatek, George Vouros, Shenghui Wang and Mikalai Yatskevich.

Finally, we also thank all the OAEI participants who made these evaluations worthwhile: Masaki Aono, John Beecher-Deighan, Boutheina Ben Yaghlane, Sadok Ben Yahia, Wayne Bethea, Jürgen Bock, Olivier Bodenreider, Gosse Bouma, Silvana Castano, Gong Cheng, Isabel F. Cruz, Carlo Curino, Jérôme David, Jean-François Djoufak-Kengue, Marc Ehrig, Daniel Engmann, Alfio Ferrara, Clayton Fink, Sandra Geisler, Jorge Gracia, Philippe Guégan, Fayçal Hamdi, Jan Hettenhausen, Bo Hu, Wei Hu, Yves R. Jean-Mary, Ningsheng Jian, Mansur R. Kabuka, Yannis Kalfoglou, Vangelis Karkaletsis, Ulas C. Keles, David Kensche, Ching-Chieh Kiu, Konstantinos Kotis, Najoua Laamari, Patrick Lambrix, Chien-Sing Lee, Dan Li, Juanzi Li, Xiang Li, Yi Li, Peng Liu, Qiang Liu, Angela Maduko, Ming Mao, Sabine Massmann, Eduardo Mena, Engelbert Mephu-Nguifo, Gianpaolo Messa, Enrico Motta, Miklos Nagy, Slawomir Niedbala, Nobal Niraula, Giorgio Orsi, Roelant Ossewaarde, Flavio Palandri-Antonelli, Jeff Z. Pan, Yefei Peng, Yuzhong Qu, Christoph Quix, Erhard Rahm, Nataliya Rassadko, Quentin Reul, Chantal Reynaud, Marta Sabou, Brigitte Safar, Hanif Seddiqui, Feng Shi, E. Patrick Shironoshita, Yahya

Slimani, Vassilis Spiliopoulos, Piotr Stolarski, Umberto Straccia, Cosmin Stroe, Heiko Stoermer, William Sunna, York Sure, He Tan, Letizia Tanca, Jie Tang, Haijun Tao, Raphaël Troncy, Alexandros G. Valarakos, Petko Valtchev, Maria Vargas-Vera, George A. Vouros, Peng Wang, Yadong Wang, Honghan Wu, Baowen Xu, Peigang Xu, Tianyi Zang, Haifa Zargayouna, Sami Zghal, Yuanyuan Zhao, Duo Zhang, Songmao Zhang, Xiao Zhang, Dongdong Zheng, Qian Zhong and Xinyu Zhong.

References

1. Bogdan Alexe, Wang Chiew Tan, and Yannis Velegrakis. Comparing and evaluating mapping systems with STBenchmark. *VLDB Endowment (PVLDB)*, 1(2):1468–1471, 2008.
2. Bogdan Alexe, Wang Chiew Tan, and Yannis Velegrakis. Stbenchmark: towards a benchmark for mapping systems. *VLDB Endowment (PVLDB)*, 1(1):230–244, 2008.
3. David Aumueller, Hong-Hai Do, Sabine Massmann, and Erhard Rahm. Schema and ontology matching with COMA++. In *Proceedings of the 24th International Conference on Management of Data (SIGMOD), Software Demonstration*, pages 906–908, Baltimore (MD US), June 2005.
4. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.
5. Philip Bernstein, Alon Halevy, and Rachel Pottinger. A vision of management of complex models. *ACM SIGMOD Record*, 29(4):55–63, 2000.
6. Christian Bizer and Andreas Schultz. The berlin SPARQL benchmark. *International Journal of Semantic Web and Information Systems*, 5(2):1–24, 2009.
7. Paolo Bouquet, Marc Ehrig, Jérôme Euzenat, Enrico Franconi, Pascal Hitzler, Markus Krotzsch, Luciano Serafini, Giorgios Stamou, York Sure, and Sergio Tessaris. Specification of a common framework for characterizing alignment. Deliverable D2.2.1v2, Knowledge web NoE, December 2004.
8. Caterina Caracciolo, Jérôme Euzenat, Laura Hollink, Ryutaro Ichise, Antoine Isaac, Véronique Malaisé, Christian Meilicke, Juan Pane, Pavel Shvaiko, Heiner Stuckenschmidt, Ondřej Šváb Zamazal, and Vojtech Svatek. Results of the ontology alignment evaluation initiative 2008. In *Proceedings of the ISWC 2008 Workshop on Ontology Matching*, pages 73–119, Karlsruhe (DE), October 2007.
9. Raúl García Castro, Diana Maynard, Doug Foxvog, Holger Wache, and Rafael González-Cabero. Specification of a methodology, general criteria, and benchmark suites for benchmarking ontology tools. Deliverable D2.1.4, Knowledge web NoE, February 2004.
10. Isabel Cruz, Flavio Palandri Antonelli, and Cosmin Stroe. Agreementmaker: efficient matching for large real-world schemas and ontologies. *VLDB Endowment*, 2(2):1586–1589, 2009.
11. Isabel Cruz, Flavio Palandri Antonelli, Cosmin Stroe, Ulas C. Keles, and Angela Maduko. Using AgreementMaker to align ontologies for OAEI 2009: Overview, results, and outlook. In *Proceedings of the ISWC 2009 Workshop on Ontology Matching*, pages 135–146, October 2009.
12. Isabel F. Cruz, Flavio Palandri Antonelli, and Cosmin Stroe. Agreementmaker: Efficient matching for large real-world schemas and ontologies. *VLDB Endowment (PVLDB)*, 2(2):1586–1589, 2009.
13. Jérôme David. AROMA results for OAEI 2009. In *Proceedings of the ISWC 2009 Workshop on Ontology Matching*, pages 147–151, October 2009.
14. Jérôme David, Jérôme Euzenat, François Scharffe, and Cássia Trojahn. The Alignment API 4.0. *Semantic web journal*, 2(1), 2011.

15. Robin Dhamankar, Yoonkyong Lee, An-Hai Doan, Alon Halevy, and Pedro Domingos. iMAP: Discovering complex semantic matches between database schemas. In *Proceedings of the 23rd International Conference on Management of Data (SIGMOD)*, pages 383–394, Paris (FR), June 2004.
16. Hong-Hai Do, Sergei Melnik, and Erhard Rahm. Comparison of schema matching evaluations. In *Proceedings of the Workshop on Web, Web-Services, and Database Systems*, volume 2593 of *Lecture notes in computer science*, pages 221–237, Erfurt (DE), October 2002.
17. Fabien Duchateau and Zohra Bellahsene. Measuring the quality of an integrated schema. In *Proceedings of the 29th International Conference on Conceptual Modeling (ER)*, pages 261–273, Vancouver (CA), November 2010.
18. Fabien Duchateau, Zohra Bellahsene, and Ela Hunt. XBenchMatch: a benchmark for xml schema matching tools. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB)*, pages 1318–1321, Vienna (AT), September 2007.
19. Fabien Duchateau, Remi Coletta, Zohra Bellahsene, and Renée J. Miller. (not) yet another matcher. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1537–1540, Hong Kong (CN), November 2009.
20. Marc Ehrig and Jérôme Euzenat. Relaxed precision and recall for ontology matching. In Benjamin Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors, *Proceedings of the Workshop on Integrating Ontologies*, volume 156, page 8, August 2005.
21. Jérôme Euzenat. An API for ontology alignment. In *Proceedings of the 3rd International Semantic Web Conference (ISWC)*, volume 3298 of *Lecture Notes in Computer Science*, pages 698–712, Hiroshima (JP), November 2004.
22. Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 348–353, Hyderabad (IN), January 2007.
23. Jérôme Euzenat, Alfio Ferrara, Laura Hollink, Antoine Isaac, Cliff Joslyn, Véronique Malaisé, Christian Meilicke, Andriy Nikolov, Juan Pane, Marta Sabou, François Scharffe, Pavel Shvaiko, Vassilkis Spiliopoulos, Heiner Stuckenschmidt, Ondřej Šváb Zamazal, Vojtech Svatek, Cassia Trojahn, George Vouros, and Shenghui Wang. Results of the ontology alignment evaluation initiative 2009. In *Proceedings of the ISWC 2009 Workshop on Ontology Matching*, pages 73–126, Washington (DC US), October 2009.
24. Jérôme Euzenat, Alfio Ferrara, Christian Meilicke, Juan Pane, François Scharffe, Pavel Shvaiko, Heiner Stuckenschmidt, Ondřej Šváb Zamazal, Vojtech Svatek, and Cassia Trojahn. Results of the ontology alignment evaluation initiative 2010. In *Proceedings of the ISWC 2010 Workshop on Ontology Matching*, pages 85–125, Shanghai, China, 2010.
25. Jérôme Euzenat, Antoine Isaac, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Ondřej Šváb, Vojtech Svatek, Willem Robert Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative 2007. In *Proceedings of the ISWC 2007 Workshop on Ontology Matching*, pages 96–132, Busan (KR), November 2007.
26. Jérôme Euzenat, Malgorzata Mochol, Pavel Shvaiko, Heiner Stuckenschmidt, Ondřej Šváb, Vojtech Svatek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative 2006. In *Proceedings of the ISWC 2006 Workshop on Ontology Matching*, pages 73–95, Athens (GA US), November 2006.
27. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer, Heidelberg (DE), 2007.
28. Jérôme Euzenat, Heiner Stuckenschmidt, and Mikalai Yatskevich. Introduction to the ontology alignment evaluation 2005. In *Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies*, Banff (CA), October 2005.
29. Raul Garcia-Castro, Asunción Gómez-Pérez, and Jesus Prieto-Gonzalez. IBSE: An OWL interoperability evaluation infrastructure. In Christine Golbreich, Aditya Kalyanpur, and Bijan Parsia, editors, *Proceedings of the Workshop on OWL: Experiences and Directions*

(OWLED), volume 258 of *CEUR Workshop Proceedings*, pages 1–10, Innsbruck, Austria, June 2007.
30. Fausto Giunchiglia, Mikalai Yatskevich Paolo, Avesani, and Pavel Shvaiko. A large scale dataset for the evaluation of ontology matching systems. *The Knowledge Engineering Review Journal (KER)*, 24(2):137–157, 2009.
31. Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. Semantic schema matching. In *Proceedings of the 13rd International Conference on Cooperative Information Systems (CoopIS)*, volume 3761 of *Lecture notes in computer science*, pages 347–365, Agia Napa (CY), November 2005.
32. Fausto Giunchiglia, Mikalai Yatskevich, and Pavel Shvaiko. Semantic matching: Algorithms and implementation. *Journal on Data Semantics*, IX:1–38, 2007.
33. Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. LUBM: A benchmark for OWL knowledge base systems. *Journal of Web Semantics*, 3(2):158–182, 2005.
34. Laura Hollink, Mark van Assem, Shenghui Wang, Antoine Isaac, and Guus Schreiber. Two variations on ontology alignment evaluation: Methodological issues. In *Proceedings of the 5th European Semantic Web Conference (ESWC)*, pages 388–401, Tenerife (ES), June 2008.
35. Ian Horrocks. Ontologies and the semantic web. *Communications of the ACM*, 51(11):58–67, 2008.
36. Antoine Isaac, Henk Matthezing, Lourens van der Meij, Stefan Schlobach, Shenghui Wang, and Claus Zinn. Putting ontology alignment in context: Usage scenarios, deployment and evaluation in a library case. In *Proceedings of the 5th European Semantic Web Conference (ESWC)*, pages 402–417, Tenerife (ES), June 2008.
37. Yves R. Jean-Mary, E. Patrick Shironoshita, and Mansur R. Kabuka. Ontology matching with semantic verification. *Journal of Web Semantics*, 7(3):235–251, 2009.
38. Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18(1):1–31, 2003.
39. David Kensche, Christoph Quix, Xiang Li 0002, Yong Li, and Matthias Jarke. Generic schema mappings for composition and query answering. *Data and Knowledge Engineering*, 68(7):599–621, 2009.
40. David Kensche, Christoph Quix, Mohamed Amine Chatti, and Matthias Jarke. Gerome: A generic role based metamodel for model management. *Journal on Data Semantics*, 8:82–117, 2007.
41. Ulrich Kuster and Birgitta Konig-Ries. Towards standard test collections for the empirical evaluation of semantic web service approaches. *International Journal of Semantic Computing*, 2(3):381–402, 2008.
42. Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the 21st Symposium on Principles of Database Systems (PODS)*, pages 233–246, Madison (WI US), June 2002.
43. Marko Luther, Thorsten Liebig, Sebastien Bohm, and Olaf Noppens. Who the heck is the father of bob? In *Proceedings of the 6th European Semantic Web Conference (ESWC)*, Heraklion (Greece), May–June 2009.
44. Li Ma, Yang Yang, Zhaoming Qiu, Guotong Xie, Yue Pan, and Shengping Liu. Towards a complete OWL ontology benchmark. In Y. Sure and J. Domingue, editors, *Proceedings of the 3rd European Semantic Web Conference (ESWC)*, volume 4011 of *Lecture Notes in Computer Science*, pages 125–139, Berlin, Heidelberg, June 2006.
45. Jayant Madhavan, Philip Bernstein, and Erhard Rahm. Generic schema matching with Cupid. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)*, pages 48–58, Roma (IT), September 2001.
46. Vincenzo Maltese, Fausto Giunchiglia, and Aliaksandr Autayeu. Save up to 99% of your time in mapping validation. In *OTM Conferences (2)*, volume 6427 of *Lecture Notes in Computer Science*, pages 1044–1060, Heraklion (Greece), 2010.

47. Anan Marie and Avigdor Gal. Boosting schema matchers. In *Proceedings of the 16th International Conference on Cooperative Information Systems (CoopIS)*, volume 5331 of *Lecture Notes in Computer Science*, pages 283–300, Monterrey (MX), November 2008.
48. Christian Meilicke and Heiner Stuckenschmidt. Incoherence as a basis for measuring the quality of ontology mappings. In *Proceedings of the ISWC 2008 Workshop on Ontology Matching*, pages 1–12, Karlsruhe (DE), October 2008.
49. Peter Mork, Len Seligman, Arnon Rosenthal, Joel Korb, and Chris Wolf. The Harmony Integration Workbench. *Journal on Data Semantics*, XI:65–93, 2008.
50. Miklos Nagy, Maria Vargas-Vera, and Enrico Motta. DSSim-ontology mapping with uncertainty. In *Proceedings of the ISWC 2006 Workshop on Ontology Matching*, pages 115–123, November 2006.
51. Charles Petrie, Tiziana Margaria, Holger Lausen, and Michal Zaremba, editors. *Semantic Web Services Challenge - Results from the First Year*, volume 8 of *Semantic Web and Beyond*. Springer Verlag, 2009.
52. Quentin Reul and Jeff Z. Pan. KOSIMap: ontology alignments results for OAEI 2009. In *Proceedings of the ISWC 2009 Workshop on Ontology Matching*, pages 177–185, October 2009.
53. Dominique Ritze, Christian Meilicke, Ondřej Šváb Zamazal, and Heiner Stuckenschmidt. A pattern-based ontology matching approach for detecting complex correspondences. In *Proceedings of the ISWC 2009 Workshop on Ontology Matching*, Washington DC (USA), October 2009.
54. Marta Sabou and Jorge Gracia. Spider: bringing non-equivalence mappings to OAEI. In *Proceedings of the ISWC 2008 Workshop on Ontology Matching*, pages 199–205, Karlsruhe (DE), October 2008.
55. Michael Schmidt, Thomas Hornung, Georg Lausen, and Christoph Pinkel. Sp^2bench: A SPARQL performance benchmark. In *Proceedings of the 25th International Conference on Data Engineering (ICDE)*, pages 222–233, Shanghai (China), March–April 2009. IEEE.
56. Md. Hanif Seddiqui and Masaki Aono. Anchor-Flood: results for OAEI 2009. In *Proceedings of the ISWC 2009 Workshop on Ontology Matching*, pages 127–134, October 2009.
57. Len Seligman, Peter Mork, Alon Y. Halevy, Ken Smith, Michael J. Carey, Kuang Chen, Chris Wolf, Jayant Madhavan, Akshay Kannan, and Doug Burdick. Openii: an open source information integration toolkit. In *Proceedings of the 29th International Conference on Management of Data (SIGMOD)*, pages 1057–1060, Indianapolis (IN US), 2010.
58. Inanç Seylan, Enrico Franconi, and Jos de Bruijn. Effective query rewriting with ontologies over dboxes. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 923–925, Pasadena (USA), July 2009.
59. Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics*, IV:146–171, 2005.
60. Pavel Shvaiko and Jérôme Euzenat. Ten challenges for ontology matching. In *Proceedings of the 7th International Conference on Ontologies, Databases, and Applications of Semantics*, pages 1163–1181, Monterrey (MX), November 2008.
61. Giorgos Stoilos, Bernardo Cuenca Grau, and Ian Horrocks. How incomplete is your semantic web reasoner? In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010)*, pages 11–15, Atlanta (USA), July 2010. AAAI Press.
62. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the 3rd ISWC Workshop on Evaluation of Ontology-based tools (EON)*, Hiroshima (JP), November 2004.
63. York Sure, Asuncion Gómez-Pérez, Walter Daelemans, Marie-Laure Reinberger, Nicola Guarino, and Natalya Noy. Why evaluate ontology technologies? because it works! *IEEE Intelligent Systems*, 19(4):74–81, 2004.

64. Cassia Trojahn, Christian Meilicke, Jérôme Euzenat, and Heiner Stuckenschmidt. Automating oaei campaigns (first report). In *Proceedings of the International Workshop on Evaluation of Semantic Technologies (IWEST)*, 2010.
65. Giovanni Tummarello, Renaud Delbru, and Eyal Oren. Sindice.com: Weaving the open linked data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, pages 552–565, Busan (KR), November 2007.
66. Lorenzino Vaccari, Pavel Shvaiko, Juan Pane, Paolo Besana, and Maurizio Marchese. An evaluation of ontology matching in geo-service applications. *GeoInformatica*, 2011, in press.
67. Willem Robert van Hage, Sophia Katrenko, and Guus Schreiber. A method to combine linguistic ontology-mapping techniques. In *Proceedings of the 4th International Semantic Web Conference (ISWC)*, pages 732–744, Galway (IE), November 2005.
68. Willem Robert van Hage, Hap Kolb, and Guus Schreiber. Relevance-based Evaluation of Alignment Approaches: The OAEI 2007 Food Task Revisited. In *Proceedings of the ISWC 2008 Workshop on Ontology Matching*, pages 234–238, October 2008.
69. Ondřej Šváb, Vojtěch Svatek, Petr Berka, Dušan Rak, and Petr Tomášek. Ontofarm: Towards an experimental collection of parallel ontologies. In *Proceedings of the 4th International Semantic Web Conference (ISWC) – Poster Track*, pages 1–3, Galway (IE), November 2005.
70. Antoine Zimmermann and Jérôme Euzenat. Three semantics for distributed systems and their relations with alignment composition. In *Proceedings of the 5th International Semantic Web Conference (ISWC)*, volume 4273 of *Lecture notes in computer science*, pages 16–29, Athens (GA US), October 2006.