Results of the Ontology Alignment Evaluation Initiative 2017*

Manel Achichi[1], Michelle Cheatham[2], Zlatan Dragisic[3], Jérôme Euzenat[4],
Daniel Faria[5], Alfio Ferrara[6], Giorgos Flouris[7], Irini Fundulaki[7], Ian Harrow[8],
Valentina Ivanova[3], Ernesto Jiménez-Ruiz[9], Kristian Kolthoff[10], Elena Kuss[10],
Patrick Lambrix[3], Henrik Leopold[11], Huanyu Li[3], Christian Meilicke[10],
Majid Mohammadi[12], Stefano Montanelli[6], Catia Pesquita[13], Tzanina Saveta[7],
Pavel Shvaiko[14], Andrea Splendiani[8], Heiner Stuckenschmidt[10], Elodie Thiéblin[15],
Konstantin Todorov[1], Cássia Trojahn[15], and Ondřej Zamazal[16]

[1] LIRMM/University of Montpellier, France
lastname@lirmm.fr
[2] Data Semantics (DaSe) Laboratory, Wright State University, USA
michelle.cheatham@wright.edu
[3] Linköping University & Swedish e-Science Research Center, Linköping, Sweden
{zlatan.dragisic,valentina.ivanova,patrick.lambrix,huanyu.li}@liu.se
[4] INRIA & Univ. Grenoble Alpes, Grenoble, France
Jerome.Euzenat@inria.fr
[5] Instituto Gulbenkian de Ciência, Lisbon, Portugal
dfaria@igc.gulbenkian.pt
[6] Università degli studi di Milano, Italy
{alfio.ferrara,stefano.montanelli}@unimi.it
[7] Institute of Computer Science-FORTH, Heraklion, Greece
{jsaveta,fgeo,fundul}@ics.forth.gr
[8] Pistoia Alliance Inc., USA
{ian.harrow,andrea.splendiani}@pistoiaalliance.org
[9] Department of Informatics, University of Oslo, Norway
ernestoj@ifi.uio.no
[10] University of Mannheim, Germany
{christian,elena,heiner}@informatik.uni-mannheim.de
[11] Vrije Universiteit Amsterdam, Netherlands
h.leopold@vu.nl
[12] Faculty of Technology, Policy, and Management, Technical University of Delft, Netherlands
m.mohammadi@tudelft.nl
[13] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
cpesquita@di.fc.ul.pt
[14] TasLab, Informatica Trentina, Trento, Italy
pavel.shvaiko@infotn.it
[15] IRIT & Université Toulouse II, Toulouse, France
{cassia.trojahn}@irit.fr
[16] University of Economics, Prague, Czech Republic
ondrej.zamazal@vse.cz

Abstract. Ontology matching consists of finding correspondences between semantically related entities of different ontologies.

[*] Note that the only official results of the campaign are on the OAEI web site.

The Ontology Alignment Evaluation Initiative (OAEI) aims at comparing ontology matching systems on precisely defined test cases. These test cases can be based on ontologies of different levels of complexity (from simple thesauri to expressive OWL ontologies) and use different evaluation modalities (e.g., blind evaluation, open evaluation, or consensus). The OAEI 2017 campaign offered 9 tracks with 23 test cases, and was attended by 21 participants. This paper is an overall presentation of that campaign.

1 Introduction

The Ontology Alignment Evaluation Initiative[1] (OAEI) is a coordinated international initiative, which organizes the evaluation of an increasing number of ontology matching systems [20, 22]. The main goal of the OAEI is to compare systems and algorithms openly and on the same basis, in order to allow anyone to draw conclusions about the best matching strategies. Furthermore, our ambition is that, from such evaluations, tool developers can improve their systems.

Two first events were organized in 2004: *(i)* the Information Interpretation and Integration Conference (I3CON) held at the NIST Performance Metrics for Intelligent Systems (PerMIS) workshop and *(ii)* the Ontology Alignment Contest held at the Evaluation of Ontology-based Tools (EON) workshop of the annual International Semantic Web Conference (ISWC) [46]. Then, a unique OAEI campaign occurred in 2005 at the workshop on Integrating Ontologies held in conjunction with the International Conference on Knowledge Capture (K-Cap) [5]. From 2006 until the present, the OAEI campaigns were held at the Ontology Matching workshop, collocated with ISWC [2, 3, 7–9, 13, 16–19, 21], which this year took place in Vienna, Austria[2].

Since 2011, we have been using an environment for automatically processing evaluations (§2.2) which was developed within the SEALS (Semantic Evaluation At Large Scale) project[3]. SEALS provided a software infrastructure for automatically executing evaluations and evaluation campaigns for typical semantic web tools, including ontology matching. In the OAEI 2017, a novel evaluation environment called HOBBIT (§10) was adopted for the novel HOBBIT Link Discovery track. Except for this track, all systems were executed under the SEALS client in all other tracks. The Benchmark track was discontinued in this edition of the OAEI.

This paper synthesizes the 2017 evaluation campaign and introduces the results provided in the papers of the participants. The remainder of the paper is organised as follows: in Section 2, we present the overall evaluation methodology that has been used; Sections 3-11 discuss the settings and the results of each of the test cases; Section 13 overviews lessons learned from the campaign; and finally, Section 14 concludes the paper.

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

[2] `http://om2017.ontologymatching.org`

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

2 General methodology

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

2.1 Tracks and test cases

This year's OAEI campaign consisted of 9 tracks gathering 23 test cases, and different evaluation modalities:

Expressive Ontology tracks offer alignments between real world ontologies expressed in OWL:

- **Anatomy (§3):** The anatomy track comprises a single test case consisting of matching the Adult Mouse Anatomy (2744 classes) and a small fragment of the NCI Thesaurus (3304 classes) describing the human anatomy. Results are evaluated automatically against a manually curated reference alignment.
- **Conference (§4):** The conference track comprises a single test case that is a suite of 21 matching tasks corresponding to the pairwise combination of 7 ontologies describing the domain of organizing conferences. Results are evaluated automatically against reference alignments in several modalities, and by using logical reasoning techniques.
- **Large biomedical ontologies (§5):** The largebio track comprises 6 test cases involving 3 large and semantically rich biomedical ontologies: FMA, SNOMED-CT, and NCI Thesaurus. These test cases correspond to the pairwise combination of these ontologies in two variants: small overlapping fragments, in which only overlapping sections of the ontologies are matched, and whole ontologies. The evaluation is based on reference alignments automatically derived from the UMLS Metathesaurus, with mappings causing logical incoherence flagged so as not to be taken into account.
- **Disease & Phenotype (§6):** The disease & phenotype track comprises 4 test cases that involve 6 biomedical ontologies covering the disease and phenotype domains: HPO versus MP, DOID versus ORDO, HPO versus MeSH, and HPO versus OMIM. The evaluation has been performed according to (1) a consensus alignment generated from those produced by the participating systems, (2) a set of manually generated mappings, and (3) a manual assessment of unique mappings (i.e., mappings that are not suggested by other systems).

Multilingual tracks offer alignments between ontologies in different languages:

- **Multifarm (§7):** The multifarm track is based on a subset of the Conference data set translated into ten different languages, in addition to their original English: Arabic, Chinese, Czech, Dutch, French, German, Italian, Portuguese, Russian, and Spanish. It consists of two test cases: same ontologies, where two versions of the same ontology in different languages are matched, and different ontologies, in which two different ontologies in different languages are matched. In total, 45 language pairings are evaluated, meaning that the same ontologies test case comprises 315 matching tasks, and the different ontologies test case comprises 945 matching tasks. Results are evaluated automatically against reference alignments.

Interactive tracks provide simulated user interaction to enable the benchmarking of algorithms designed to make use of it, with respect to both the improvement in the results and the workload of the user:

- **Interactive Matching Evaluation (§8):** The Interactive track is based on the test cases from the anatomy and conference tracks. An Oracle, which matching tools can access programmatically, simulates user feedback by querying the reference alignment of the test case. The Oracle can generate erroneous responses at a given rate, to simulate user errors. The evaluation is based on the same reference alignments, and contemplates the number of user interactions and the fraction of erroneous responses received by the tool, in addition to the standard evaluation parameters.

Instance Matching tracks focus on alignments between ontology instances expressed in the form of OWL Aboxes:

- **Instance Matching (§9).** The instance track comprises two independent sub-tracks:
  - **SYNTHETIC:** This sub-track consists of matching instances that are found to refer to the same real-world entity corresponding to a creative work (that can be a news item, blog post or programme). It includes two evaluation modalities, *Sandbox* and *Mainbox*, which differ on the number of instances to match. The evaluation is automatic, based on a reference alignment, and partially blind – matching tools have access only to the *Sandbox* reference alignment.
  - **DOREMUS:** This sub-track consists of matching real world datasets about classical music artworks from two major French cultural institutions: the French National Library (BnF) and the Philharmonie de Paris (PP). Both datasets use the same vocabulary, the DOREMUS model, issued from the DOREMUS project[4]. This sub-track comprises two different test cases called *heterogeneities* (HT) and *false-positives trap* (FPT) characterized by different degrees of heterogeneity in artwork descriptions. The evaluation is automatic and based on reference alignments.
- **HOBBIT Link Discovery (§10).** The HOBBIT track aims to deal with link discovery for spatial data represented as trajectories or traces i.e., sequences of longitude, latitude pairs. It comprises two test cases: Linking and Spatial. The Linking test case consists in matching traces that have been modified using string-based approaches, different date and coordinate formats, and by addition and/or deletion of intermediate points. In the Spatial test case, the goal is to identify DE-9IM (Dimensionally Extended nine-Intersection Model) topological relations between traces: *Equals*, *Disjoint*, *Touches*, *Contains/Within*, *Covers/CoveredBy*, *Intersects*, *Crosses*, *Overlaps*. For each relation, a different pair of source and target datasets is given to the participants, so the test

[4] http://www.doremus.org

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

test	formalism	relations	confidence	modalities	language	SEALS
anatomy	OWL	=	[0 1]	open	EN	√
conference	OWL	=, <=	[0 1]	open+blind	EN	√
largebio	OWL	=	[0 1]	open	EN	√
phenotype	OWL	=	[0 1]	blind	EN	√
multifarm	OWL	=	[0 1]	open+blind	AR, CZ, CN, DE, EN, ES, FR, IT, NL, RU, PT	√
interactive	OWL	=, <=	[0 1]	open	EN	√
instance	OWL	=	[0 1]	open+blind	EN	√
HOBBIT	OWL	=, spatial	N/A	open+blind	EN, N/A	
process model	OWL	<=	[0 1]	open+blind	EN	√

case consists of 8 individual matching tasks. In both test cases, two evaluation modalities, *Sandbox* and *Mainbox*, were considered, differing on the number of instances to match. The evaluation is automatic and based on reference alignments.

Process Model Matching (§11): The process model track is concerned with the application of ontology matching techniques to the problem of matching process models. It comprises two test cases used in the Process Model Matching Campaign 2015 [4] which have been converted to an ontological representation, with process model entities being represented as ontology instances. The first test case contains nine process models which represent the application process for a master program of German universities as well as reference alignments between all pairs of models. The second test case consists of process models which describe the process of registering a newborn child in different countries. The evaluation is automatic, based on reference alignments, and uses standard precision and recall measures as well as a probabilistic variant described in [29].

Table 1 summarizes the variation in the proposed test cases.

2.2 The SEALS client

Since 2011, tool developers had to implement a simple interface and to wrap their tools in a predefined way including all required libraries and resources. A tutorial for tool wrapping was provided to the participants, describing how to wrap a tool and how to use the SEALS client to run a full evaluation locally. This client is then executed by the track organizers to run the evaluation. This approach ensures the reproducibility and comparability of the results of all systems.

2.3 Preparatory phase

Ontologies to be matched and (where applicable) reference alignments have been provided in advance during the period between June 1^{st} and July 15^{th}, 2017. This gave potential participants the occasion to send observations, bug corrections, remarks and other test cases to the organizers. The goal of this preparatory period is to ensure that the delivered tests make sense to the participants. The final test base was released on July 15^{th}, 2017 and did not evolve after that.

2.4 Execution phase

During the execution phase, participants used their systems to automatically match the test case ontologies. In most cases, ontologies are described in OWL-DL and serialized in the RDF/XML format [11]. Participants can self-evaluate their results either by comparing their output with reference alignments or by using the SEALS client to compute precision and recall. They can tune their systems with respect to the non blind evaluation as long as the rules published on the OAEI web site are satisfied. This phase has been conducted between July 15^{th} and August 31^{st}, 2017, except for the HOBBIT track which was extended until September 15^{th}, 2017. Like last year, we requested a mandatory registration of systems and a preliminary evaluation of wrapped systems by July 31st, to alleviate the burden of debugging systems with respect to issues with the SEALS client during the Evaluation phase.

2.5 Evaluation phase

Participants were required to submit their SEALS-wrapped tools by August 31^{st}, 2017, and their HOBBIT-wrapped tool by September 15^{th}, 2017. Tools were then tested by the organizers and minor problems were reported to some tool developers, who were given the opportunity to fix their tools and resubmit them.

Initial results were provided directly to the participants between September 1^{st} and October 15^{th}, 2017. The final results for most tracks were published on the respective pages of the OAEI website by October 15^{th}, although some tracks were delayed.

The standard evaluation measures are precision, recall and F-measure computed against the reference alignments. More details on the evaluation are given in the sections for the test cases.

2.6 Comments on the execution

Following an initial period of growth, the number of OAEI participants has remained approximately constant since 2012, at slightly over 20 (see Figure 1). This year was no exception, as we counted 21 participating systems. Table 2 lists the participants and the tracks in which they competed. Some matching systems participated with different variants (DiSMatch and LogMap) whereas others were evaluated with different configurations, as requested by developers (see test case sections for details).

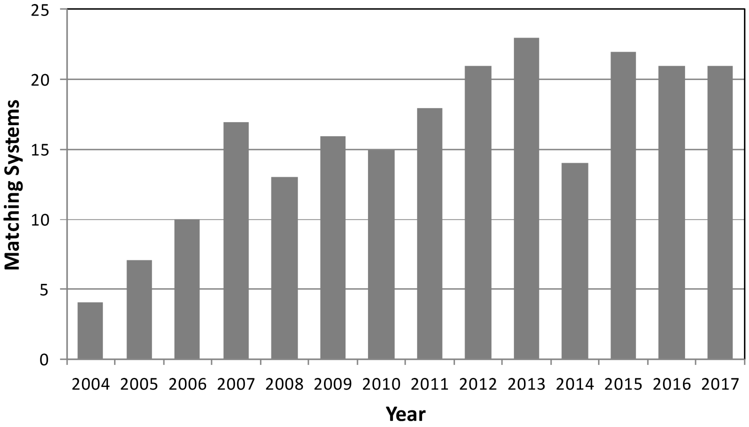


Fig. 1. Number of participating systems per year in the OAEI.

Table 2. Participants and the status of their submissions.

System	ALIN	AML	CroLOM	DiSMatch-ar	DiSMatch-sg	DiSMatch-tr	I-Match	KEPLER	Legato	LogMap	LogMap-Bio	LogMapLt	njuLink	ONTMAT	POMap	RADON	SANOM	Silk	Wiki2	XMap	YAM-BIO	Total=21
Confidence	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	-	✓	-	✓	✓	✓	✓	-	✓	-	16
anatomy	●	●	○	○	○	○	○	●	○	●	●	●	○	○	●	○	●	○	●	●	●	11
conference	●	●	○	○	○	○	○	●	○	●	○	●	○	●	●	○	●	○	●	●	○	10
largebio	○	●	○	○	○	○	○	◐	○	●	●	●	○	○	◐	○	◐	○	◐	●	●	10
phenotype	○	●	○	●	●	●	○	◐	○	●	●	◐	○	○	◐	○	○	○	○	◐	◐	11
multifarm	○	●	●	○	○	○	○	●	○	●	○	○	○	○	○	○	◐	○	●	◐	○	7
interactive	●	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	●	○	4
process model	○	●	○	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	3
instance	○	●	○	○	○	○	●	○	●	●	○	○	◐	○	○	○	○	○	○	○	○	5
hobbit ld	○	●	○	○	○	○	●	○	○	○	○	○	○	○	○	◐	○	◐	○	○	○	4
total	3	9	1	1	1	1	3	5	1	8	3	4	1	1	4	1	4	1	4	6	3	65

Confidence pertains to the confidance scores returned by the system, with ✓ indicating that they are non-boolean; ○ indicates that the system did not participate in the track; ● indicates that it participated fully in the track; and ◐ indicates that it participated in or completed only part of the tasks of the track.

3 Anatomy

The anatomy test case confronts matching systems with two fragments of biomedical ontologies which describe the human anatomy[5] and the anatomy of the mouse[6]. This data set has been used since 2007 with some improvements over the years [15].

[5] http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/

3.1 Experimental Setting

We conducted experiments by executing each system in its standard setting and we compare precision, recall, F-measure and recall+ against a manually curated reference alignment. Recall+ indicates the amount of detected non-trivial correspondences, i.e., correspondence that do not have the same normalized label. The approach that generates only trivial correspondences is depicted as baseline StringEquiv in the following section.

We ran the systems on a server with 3.46 GHz (6 cores) and 8GB allocated RAM, using the SEALS client. However, we changed the way precision and recall are computed by removing trivial correspondences in the oboInOwl namespace like:

```
http://...oboInOwl#Synonym = http://...oboInOwl#Synonym
```

as well as correspondences expressing relations different from equivalence. Thus, the results generated by the SEALS client vary in some cases by 0.5% compared to the results presented below. Using the Pellet reasoner we also checked whether the generated alignment is coherent, i.e., that there are no unsatisfiable classes when the ontologies are merged with the alignment.

3.2 Results

In Table 3, we show the results of the 11 participating systems that generated an alignment, including 3 versions of LogMap. A number of systems participated in the anatomy track for the first time this year: KEPLER, POMap, SANOM, WikiV2, and YAM-BIO. For more details, we refer the reader to the papers presenting the systems.

Table 3. Comparison, ordered by F-measure, against the reference alignment, runtime is measured in seconds, the “size” column refers to the number of correspondences in the generated alignment.

Matcher	Runtime	Size	Precision	F-measure	Recall	Recall+	Coherent
AML	47	1493	0.95	0.943	0.936	0.832	√
YAM-BIO	70	1474	0.948	0.935	0.922	0.794	-
POMap	808	1492	0.94	0.933	0.925	0.824	-
LogMapBio	820	1534	0.889	0.894	0.899	0.733	√
XMap	37	1412	0.926	0.893	0.863	0.639	√
LogMap	22	1397	0.918	0.88	0.846	0.593	√
KEPLER	234	1173	0.958	0.836	0.741	0.316	-
LogMapLite	19	1148	0.962	0.829	0.728	0.29	-
SANOM	295	1304	0.895	0.828	0.77	0.419	-
Wiki2	2204	1260	0.883	0.802	0.734	0.356	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
ALIN	836	516	0.996	0.506	0.339	0.0	√

This year 5 out of 11 systems were able to achieve the alignment task in less than 100 seconds: LogMapLite, LogMap, XMap, AML and YAM-BIO. In 2016 and 2015, there

[6] http://www.informatics.jax.org/searches/AMA_form.shtml

were 4 out of 13 systems and 6 out of 15 systems respectively that generated an alignment in this time frame. As in the last 5 years LogMapLite has the shortest runtime. The table shows that there is no correlation between the quality of the generated alignment in terms of precision and recall and the runtime. This result had also been observed in previous OAEI campaigns.

The table also shows the results for F-measure, recall+ and the size of alignments. Regarding F-measure, the top 3 ranked systems AML, YAM-BIO, POMap achieve on F-measure above 0.93. Among these, AML achieved the highest F-measure (0.943). All of the long-term participants in the track showed comparable results in terms of F-measure to their last year's results and at least as good as the results of the best systems in OAEI 2007-2010. Regarding recall+, AML, LogMap, LogMapLite showed similar results to previous years. LogMapBio has a slight increase from 0.728 in 2016 to 0.733 in 2017. XMap decreases a bit from 0.647 to 0.639. Two new participants obtained good results for recall+, POMap scored 0.824 (second place) followed by YAM-BIO with 0.794 (third place). In terms of the number of correspondences, long-term participants computed similar numbers of correspondences as last year. AML and LogMap generated the same number of correspondences, LogMapBio generated 3 more correspondences, LogMapLite generated 1 more, ALIN generated 6 more and XMap generated 1 less.

This year, 10 out of 11 systems achieved an F-measure higher than the baseline. This is a slightly better result than last year when 9 out of 13 surpassed the baseline. Five systems produced coherent alignments, which is comparable to the last two years when 7 out of 13 and 5 out of 10 systems achieved this. Two of the three best systems with respect to F-measure (YAM-BIO and POMap) produced incoherent alignments.

3.3 Conclusions

The number of systems participating in the anatomy track has varied throughout the years. This year, it is lower than in the two previous editions, but higher than in 2014. As noted previously there are newly-joined systems as well as long-term participants.

The systems that participated in the previous edition in 2016 scored similarly to their previous results. As last year, the AML system set the top result for anatomy track with respect to F-measure. Two of the newly-joined systems (YAM-BIO and POMap) achieved 2nd and 3rd best score in terms of F-measure.

4 Conference

The conference test cases require matching several moderately expressive ontologies from the conference organisation domain.

4.1 Test data

The data set consists of 16 ontologies in the domain of organising conferences. These ontologies were developed within the OntoFarm project[7].

The main features of this test case are:

[7] http://owl.vse.cz:8080/ontofarm/

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

Ontologies differ in their numbers of classes and properties, in expressivity, but also in underlying resources.

4.2 Results

We performed three kinds of evaluations. First, we provide results in terms of F-measure, comparison with baseline matchers and results of matchers from previous OAEI editions and precision/recall triangular graph based on sharp reference alignments. Second, we provide an evaluation based on the uncertain version of the reference alignment, and finally we also provide an evaluation based on violations of consistency and conservativity principles.

Evaluation based on sharp reference alignments We evaluated the results of participants against blind reference alignments (labelled as *rar2*).[8] This includes all pairwise combinations between 7 different ontologies, i.e., 21 alignments.

We have prepared the reference alignments in two steps. First, we have generated them as a transitive closure computed on the original reference alignments. In order to obtain a coherent result, conflicting correspondences, i.e., those causing unsatisfiability, have been manually inspected and incoherency has been resolved by evaluators. The resulting reference alignments are labelled as *ra2*. Second, we detected violations of conservativity using the approach from [44] and resolved them by an evaluator. The resulting reference alignments are labelled as *rar2*. As a result, the degree of correctness and completeness of the new reference alignments is probably slightly better than for the old one. However, the differences are relatively limited. Whereas the new reference alignments are not open, the old reference alignments (labeled as *ra1* on the conference web page) are available. These represent close approximations of the new ones.

Table 4 shows the results of all participants with regard to the reference alignment *rar2*. $F_{0.5}$-measure, F_1-measure and F_2-measure are computed for the threshold that provides the optimal F_1-measure. F_1 is the harmonic mean of precision and recall where both are equally weighted; F_2 weights recall higher than precision and $F_{0.5}$ weights precision higher than recall. The matchers shown in the table are ordered according to their highest average F_1-measure. We employed two baseline matchers. edna (string edit distance matcher) was used within the benchmark test cases in previous years and with regard to performance it is very similar as the previously used *baseline2* in the conference track; StringEquiv is used within the anatomy test case. This year these baselines divide matchers into two performance groups.

[8] More details about evaluation applying other sharp reference alignments are available at the conference web page.

Table 4. The highest average $F_{[0.5|1|2]}$-measure and their corresponding precision and recall for each matcher with its F_1-optimal threshold (ordered by F_1-measure). Inc.Align. means number of incoherent alignments. Conser.V. means total number of all conservativity principle violations. Consist.V. means total number of all consistency principle violations.

Matcher	Prec.	$F_{0.5}$-m.	F_1-m.	F_2-m.	Rec.	Inc.Align.	Conser.V.	Consist.V.
AML	0.78	0.74	0.69	0.65	0.62	0	39	0
LogMap	0.77	0.72	0.66	0.6	0.57	0	25	0
XMap	0.78	0.72	0.65	0.58	0.55	1	22	4
LogMapLt	0.68	0.62	0.56	0.5	0.47	5	96	25
edna	*0.74*	*0.66*	*0.56*	*0.49*	*0.45*			
KEPLER	0.67	0.61	0.55	0.49	0.46	12	123	159
WikiV3	0.63	0.59	0.54	0.5	0.47	10	125	58
StringEquiv	*0.76*	*0.65*	*0.53*	*0.45*	*0.41*			
POMap	0.69	0.59	0.49	0.42	0.38	0	1	0
ALIN	0.86	0.6	0.41	0.31	0.27	0	0	0
SANOM	0.8	0.56	0.38	0.29	0.25	1	11	18
ONTMAT	0.06	0.07	0.1	0.19	0.41	0	1	0

With regard to the two baselines, we can group tools according to each matcher's position. In all, four tools outperformed both baselines (AML, LogMap, XMap and LogMapLt), and two newcomers (KEPLER and WikiV3) performed better than one baseline. Other matchers (POMap, ALIN, SANOM and ONTMAT) performed worse than both baselines. Four tools (ALIN, POMap, ONTMAT and SANOM) did not match properties at all. Of course, this had a negative effect on those tools' overall performance. More details about evaluation considering only classes or properties are on the conference web page. The performance of all matchers (except ONTMAT) regarding their precision, recall and F_1-measure is visualised in Figure 2. Matchers are represented as squares or triangles. Baselines are represented as circles.

Comparison with previous years with regard to rar2 Four matchers, top-performers, also participated in the Conference test cases in OAEI 2016. None of them improved with regard to F1-measure evaluation.

Evaluation based on uncertain version of reference alignments The confidence values of all matches in the sharp reference alignments for the conference track are all 1.0. For the uncertain version of this track, the confidence value of a match has been set equal to the percentage of a group of people who agreed with the match in question (this uncertain version is based on the reference alignment labeled *ra1*). One key thing to note is that the group was only asked to validate matches that were already present in the existing reference alignments – so some matches had their confidence value reduced from 1.0 to a number near 0, but no new match was added.

There are two ways that we can evaluate matchers according to these "uncertain" reference alignments, which we refer to as *discrete* and *continuous*. The discrete evaluation considers any match in the reference alignment with a confidence value of 0.5 or greater to be fully correct and those with a confidence less than 0.5 to be fully incorrect. Similarly, a matcher's match is considered a "yes" if the confidence value is greater than or equal to the matcher's threshold and a "no" otherwise. In essence, this is the same as the "sharp" evaluation approach, except that some matches have been removed because less than half of the crowdsourcing group agreed with them. The continuous evaluation strategy penalises a matcher more if it misses a match on

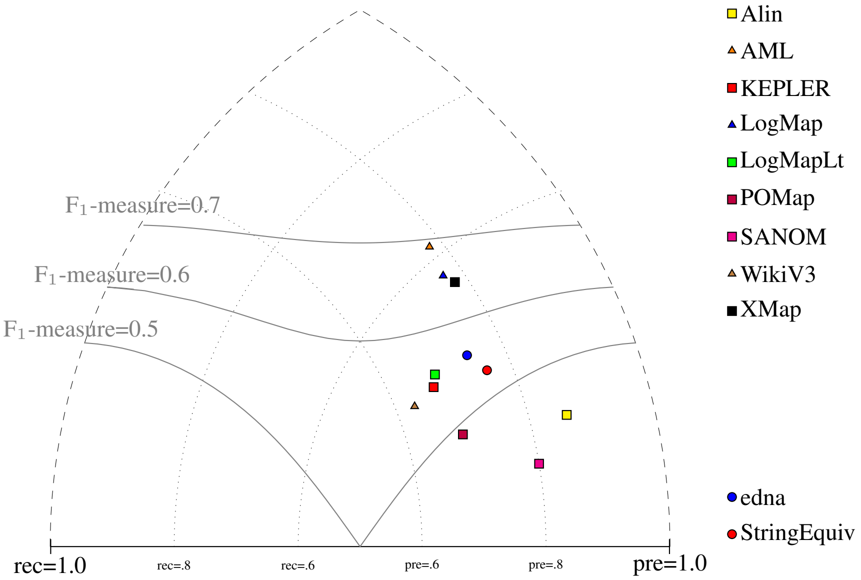


Fig. 2. Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F$_1$-measure are depicted by areas bordered by corresponding lines F$_1$-measure=0.[5|6|7].

which most people agree than if it misses a more controversial match. For instance, if $A \equiv B$ with a confidence of 0.85 in the reference alignment and a matcher gives that correspondence a confidence of 0.40, then that is counted as $0.85 \times 0.40 = 0.34$ true positive and $0.85 - 0.40 = 0.45$ false negative.

Out of the ten alignment matchers, three (ALIN, LogMapLt and ONTMAT) use 1.0 as the confidence value for all matches they identify. Two more have a narrow range of confidence values (POMap's values vary between 0.8 and 1.0, with the majority falling between 0.93 and 1.0 while SANOM's values are relatively tightly clustered between 0.73 and 0.9). The remaining five systems (AML, KEPLER, LogMap, WikiV3 and XMap) have a wide variation of confidence values.

When comparing the performance of the matchers on the uncertain reference alignments versus that on the sharp version (see Table 5), we see that in the discrete case all matchers performed the same or slightly better. Improvement in F-measure ranged from 0 to 8 percentage points over the sharp reference alignment. This was driven by increased recall, which is a result of the presence of fewer "controversial" matches in the uncertain version of the reference alignment.

The performance of most matchers is very similar regardless of whether a discrete or continuous evaluation methodology is used (provided that the threshold is optimized to achieve the highest possible F-measure in the discrete case). The primary exceptions to this are KEPLER, LogMap and SANOM. These systems perform significantly worse when evaluated using the continuous version of the metrics. In the LogMap and SANOM cases, this is because the matcher assigns low confidence values to some matches in which the labels are equivalent strings, which many crowdsourcers agreed with unless there was a compelling technical reason not to. This hurts recall, but using a low threshold value in the discrete version of the evaluation metrics 'hides' this problem. In the case of KEPLER, the issue is that entities whose labels share a word in common

Table 5. F-measure, precision, and recall of the different matchers when evaluated using the sharp (*ra1*), discrete uncertain and continuous uncertain metrics.

	Sharp			Discrete			Continuous		
Matcher	Prec.	F_1-m.	Rec.	Prec.	F_1-m.	Rec.	Prec.	F_1-m.	Rec.
ALIN	0.89	0.41	0.27	0.89	0.49	0.34	0.89	0.5	0.35
AML	0.84	0.74	0.66	0.79	0.78	0.77	0.8	0.77	0.74
KEPLER	0.76	0.59	0.48	0.76	0.67	0.6	0.58	0.62	0.68
LogMap	0.82	0.69	0.59	0.78	0.73	0.68	0.8	0.67	0.57
LogMapLt	0.73	0.59	0.5	0.72	0.67	0.62	0.72	0.67	0.63
ONTMAT	0.06	0.11	0.43	0.06	0.11	0.54	0.06	0.11	0.55
POMap	0.73	0.52	0.4	0.73	0.6	0.5	0.71	0.59	0.51
SANOM	0.81	0.38	0.25	0.81	0.45	0.31	0.81	0.38	0.25
WikiV3	0.67	0.57	0.49	0.74	0.62	0.52	0.73	0.63	0.55
XMap	0.84	0.68	0.57	0.79	0.72	0.67	0.81	0.73	0.67

have fairly high confidence values, even though they are often not equivalent. For example, "Review" and "Reviewing_Event". This hurts precision in the continuous case, but is taken care of by using a high threshold value in the discrete case.

Five matchers from this year also participated last year, and thus we are able to make some comparisons over time. The F-measures of all systems essentially held constant (within one percent) when evaluated against the uncertain reference alignments. This is in contrast to last year, in which most matchers made modest gains (in the neighborhood of 1 to 6 percent) over 2015. It seems that, barring any new advances, participating matchers have reached something of a steady state on this performance metric.

Evaluation based on violations of consistency and conservativity principles We performed evaluation based on detection of conservativity and consistency violations [44, 45]. The consistency principle states that correspondences should not lead to unsatisfiable classes in the merged ontology; the conservativity principle states that correspondences should not introduce new semantic relationships between concepts from one of the input ontologies.

Table 4 shows the number of unsatisfiable TBoxes after the ontologies are merged (Inc. Align.), the total number of all conservativity principle violations within all alignments (Conser.V.) and the total number of all consistency principle violations (Consist.V.).

Five tools (ALIN, AML, LogMap, ONTMAT and POMap) have no consistency principle violation (in comparison to seven last year) and two tools (SANOM and XMap) generated only one incoherent alignment. There is one tool (ALIN) having no conservativity principle violations. Further two tools (ONTMAT and POMap) have an average of conservativity principle violations around 1. We should note that these conservativity principle violations can be "false positives" since the entailment in the aligned ontology can be correct although it was not derivable in the single input ontologies.

4.3 Conclusions

In conclusion, this year four of ten matchers performed better than both baselines on sharp reference alignments. Further, this year five matchers generated coherent alignments (against seven matchers last year and five matchers the year before). Based on the uncertain reference alignments we can conclude that all matchers perform better on the fuzzy versus sharp version of the benchmark and eight matchers have close correspondence on the continuous and discrete version, indicating good agreement with the human matchers. Finally, none of the five matchers that also participated last year improved their performance with regard to the evaluation based on the sharp or the uncertain reference alignments.

5 Large biomedical ontologies (largebio)

The largebio test case aims at finding alignments between the large and semantically rich biomedical ontologies FMA, SNOMED-CT, and NCI, which contain 78,989, 306,591 and 66,724 classes, respectively.

5.1 Test data

The test case has been split into three matching problems: FMA-NCI, FMA-SNOMED and SNOMED-NCI. Each matching problem has been further divided in 2 tasks involving differently sized fragments of the input ontologies: small overlapping fragments versus whole ontologies (FMA and NCI) or large fragments (SNOMED-CT).

The UMLS Metathesaurus [6] has been selected as the basis for reference alignments. UMLS is currently the most comprehensive effort for integrating independently-developed medical thesauri and ontologies, including FMA, SNOMED-CT, and NCI. The extraction of mapping from UMLS is detailed in [26]).

Since alignment coherence is an aspect of ontology matching that we aim to promote, in previous editions we provided coherent reference alignments by refining the UMLS mappings using the Alcomo (alignment) debugging system [32], LogMap's (alignment) repair facility [25], or both [27].

However, concerns were raised about the validity and fairness of applying automated alignment repair techniques to make reference alignments coherent [37]. It is clear that using the original (incoherent) UMLS alignments would be penalizing to ontology matching systems that perform alignment repair. However, using automatically repaired alignments would penalize systems that do not perform alignment repair and also systems that employ a repair strategy that differs from that used on the reference alignments [37].

Thus, as of the 2014 edition, we arrived at a compromise solution that should be fair to all ontology matching systems. Instead of repairing the reference alignments as normal, by removing correspondences, we flagged the *incoherence-causing correspondences* in the alignments by setting the relation to *"?"* (unknown). These "?" correspondences will neither be considered as positive nor as negative when evaluating the participating ontology matching systems, but will simply be ignored. This way, systems that do not perform alignment repair are not penalized for finding correspondences that (despite causing incoherences) may or may not be correct, and systems that do perform alignment repair are not penalized for removing such correspondences.

To ensure that this solution was as fair as possible to all alignment repair strategies, we flagged as unknown all correspondences suppressed by any of Alcomo, LogMap or AML [39], as well as all correspondences suppressed from the reference alignments of last year's edition (using Alcomo and LogMap combined). Note that, we have used the (incomplete) repair modules of the above mentioned systems.

The flagged UMLS-based reference alignment for the OAEI 2017 campaign is summarized in Table 6.

Table 6. Number of correspondences in the reference alignments of the large biomedical ontologies tasks

Reference alignment	"=" corresp.	"?" corresp.
FMA-NCI	2,686	338
FMA-SNOMED	6,026	2,982
SNOMED-NCI	17,210	1,634

5.2 Evaluation setting, participation and success

We have run the evaluation in a Ubuntu Laptop with an Intel Core i7-4600U CPU @ 2.10GHz x 4 and allocating 15Gb of RAM. Precision, Recall and F-measure have been computed with respect to the UMLS-based reference alignment. Systems have been ordered in terms of F-measure.

In the OAEI 2017 largebio track 10 out of 21 participating systems have been able to cope with at least one of the tasks of the largebio track with a 4 hours timeout. Note that we also include the results of *Tool1* (the developers withdrew the system from the campaign) as reference. 9 systems were able to complete more than one task, while 6 systems were able to complete all tasks. This is an improvement with respect to last year results where only 4 systems were able to complete all tasks

5.3 Background knowledge

Regarding the use of background knowledge, LogMap-Bio uses BioPortal as mediating ontology provider, that is, it (automatically) retrieves from BioPortal the most suitable top-10 ontologies for the matching task.

LogMap uses normalisations and spelling variants from the general (biomedical) purpose UMLS Lexicon (a different resource with respect to the UMLS Metathesaurus).

AML has three sources of background knowledge which can be used as mediators between the input ontologies: the Uber Anatomy Ontology (Uberon), the Human Disease Ontology (DOID) and the Medical Subject Headings (MeSH).

YAM-BIO uses as background knowledge a file containing mappings from the DOID and UBERON ontologies to other ontologies like FMA, NCI or SNOMED CT.

XMAP uses synonyms provided by the UMLS Metathesaurus. Note that matching systems using UMLS Metathesaurus as background knowledge will have a **notable advantage** since the largebio reference alignment is also based on the UMLS Metathesaurus.

5.4 Alignment coherence

Together with Precision, Recall, F-measure and run times we have also evaluated the coherence of alignments. We report (1) the number of unsatisfiabilities when reasoning with the input ontologies together with the computed alignments, and (2) the ratio of unsatisfiable classes with respect to the size of the union of the input ontologies.

We have used the OWL 2 reasoner HermiT [35] to compute the number of unsatisfiable classes. For the cases in which HermiT could not cope with the input ontologies and the alignments (in less than 2 hours) we have provided a lower bound on the number of unsatisfiable classes (indicated by $\geq$) using the OWL 2 EL reasoner ELK [28].

Table 7. System runtimes (in seconds) and task completion.

System	FMA-NCI		FMA-SNOMED		SNOMED-NCI		Average	#
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6		
LogMapLite	1	10	2	18	9	22	10	6
AML	44	77	109	177	669	312	231	6
LogMap	12	92	57	477	207	652	250	6
XMap	20	130	62	625	106	563	251	6
YAM-BIO	56	279	60	468	2,202	490	593	6
Tool1	65	1,650	245	2,140	481	1,150	955	6
LogMapBio	1,098	1,552	1,223	2,951	2,779	4,728	2,389	6
POMAP	595	-	1,841	-	-	-	1,218	2
SANOM	679	-	3,123	-	-	-	1,901	2
KEPLER	601	-	3,378	-	-	-	1,990	2
Wiki2	108,953	-	-	-	-	-	108,953	1
# Systems	**11**	**10**	**7**	**7**	**7**	**7**	**10,795**	**49**

In this OAEI edition, only three distinct systems have shown alignment repair facilities: AML, LogMap and its LogMap-Bio variant, and XMap (which reuses the repair techniques from Alcomo [32]). Note that only LogMap and LogMap-Bio are able to reduce to a minimum the number of unsatisfiable classes across all tasks. Missing 9 unsatisfiable classes in the worst case (whole FMA-NCI task).

Tables 8-9 (see last two columns) show that even the most precise alignment sets may lead to a huge number of unsatisfiable classes. This proves the importance of using techniques to assess the coherence of the generated alignments if they are to be used in tasks involving reasoning. We encourage ontology matching system developers to develop their own repair techniques or to use state-of-the-art techniques such as Alcomo [32], the repair module of LogMap (LogMap-Repair) [25] or the repair module of AML [39], which have worked well in practice [27, 23].

5.5 Runtimes and task completion

Table 7 shows which systems were able to complete each of the matching tasks in less than 4 hours and the required computation times. Systems have been ordered with respect to the number of completed tasks and the average time required to complete them. Times are reported in seconds.

The last column reports the number of tasks that a system could complete. For example, 7 system (including the withdrawn system *Tool1*) were able to complete all six tasks. The last row shows the number of systems that could finish each of the tasks. The tasks involving SNOMED were also harder with respect to both computation times and the number of systems that completed the tasks.

5.6 Results for the FMA-NCI matching problem

Table 8 summarizes the results for the tasks in the FMA-NCI matching problem.

XMap and YAM-BIO achieved the highest F-measure in Task 1, while XMap and AML in Task 2. Note however that the use of background knowledge based on the UMLS Metathesaurus has an important impact in the performance of XMap. The use of background knowledge led to

Table 8. Results for the FMA-NCI matching problem.

System	Time (s)	# Corresp.	Scores			Incoherence	
			Prec.	F-m.	Rec.	Unsat.	Degree
Task 1: small FMA and NCI fragments							
XMap*	20	2,649	0.98	0.94	0.90	2	0.019%
YAM-BIO	56	2,681	0.97	0.93	0.90	800	7.8%
AML	44	2,723	0.96	0.93	0.91	2	0.019%
LogMapBio	1,098	2,807	0.93	0.92	0.91	2	0.019%
LogMap	12	2,747	0.94	0.92	0.90	2	0.019%
KEPLER	601	2,506	0.96	0.89	0.83	3,707	36.1%
Average	10,193	2,550	0.95	0.89	0.84	1,238	12.0%
LogMapLite	1	2,483	0.97	0.89	0.82	2,045	19.9%
SANOM	679	2,457	0.95	0.87	0.80	1,183	11.5%
POMAP	595	2,475	0.90	0.86	0.83	3,493	34.0%
Tool1	65	2,316	0.97	0.86	0.77	1,128	11.0%
Wiki2	108,953	2,210	0.88	0.80	0.73	1,261	12.3%
Task 2: whole FMA and NCI ontologies							
XMap*	130	2,735	0.88	0.87	0.85	9	0.006%
AML	77	2,968	0.84	0.86	0.87	10	0.007%
YAM-BIO	279	3,109	0.82	0.85	0.89	11,770	8.1%
LogMap	92	2,701	0.86	0.83	0.81	9	0.006%
LogMapBio	1,552	2,913	0.82	0.83	0.83	9	0.006%
Average	541	2,994	0.80	0.81	0.83	7,389	5.1%
LogMapLite	10	3,477	0.67	0.74	0.82	26,478	18.1%
Tool1	1,650	3,056	0.69	0.71	0.74	13,442	9.2%

*Uses background knowledge based on the UMLS Metathesaurus which is the basis of the large-bio reference alignments.

an improvement in recall from LogMap-Bio over LogMap in both tasks, but this came at the cost of precision, resulting in the two variants of the system having identical F-measures.

Note that the effectiveness of the systems decreased from Task 1 to Task 2. One reason for this is that with larger ontologies there are more plausible mapping candidates, and thus it is harder to attain both a high precision and a high recall. Another reason is that the very scale of the problem constrains the matching strategies that systems can employ: AML for example, foregoes its matching algorithms that are computationally more complex when handling very large ontologies, due to efficiency concerns.

The size of Task 2 prove a problem for a number of systems, which were unable to complete it within the allotted time: POMAP, SANOM, KEPLER and Wiki2.

5.7 Results for the FMA-SNOMED matching problem

Table 9 summarizes the results for the tasks in the FMA-SNOMED matching problem.

XMap produced the best results in terms of both Recall and F-measure in Task 3 and Task 4, but again, we must highlight that it uses background knowledge based on the UMLS Metathesaurus. Among the other systems, AML and YAM-BIO achieved the highest F-measure in Tasks 3 and 4, respectively.

Table 9. Results for the FMA-SNOMED matching problem.

System	Time (s)	# Corresp.	Scores			Incoherence	
			Prec.	F-m.	Rec.	Unsat.	Degree
Task 3: small FMA and SNOMED fragments							
XMap*	62	7,400	0.97	0.91	0.85	0	0.0%
AML	109	6,988	0.92	0.84	0.76	0	0.0%
YAM-BIO	60	6,817	0.97	0.83	0.73	13,240	56.1%
LogMapBio	1,223	6,315	0.95	0.80	0.69	1	0.004%
LogMap	57	6,282	0.95	0.80	0.69	1	0.004%
Average	1,010	4,623	0.89	0.62	0.51	2,141	9.1%
KEPLER	3,378	4,005	0.82	0.56	0.42	3,335	14.1%
SANOM	3,123	3,146	0.69	0.42	0.30	2,768	11.7%
POMAP	1,841	2,655	0.68	0.42	0.30	1,013	4.3%
LogMapLite	2	1,644	0.97	0.34	0.21	771	3.3%
Tool1	245	979	0.99	0.24	0.14	287	1.2%
Task 4: whole FMA ontology with SNOMED large fragment							
XMap*	625	8,665	0.77	0.81	0.84	0	0.0%
YAM-BIO	468	7,171	0.89	0.80	0.73	54,081	26.8%
AML	177	6,571	0.88	0.77	0.69	0	0.0%
LogMap	477	6,394	0.84	0.73	0.65	0	0.0%
LogMapBio	2,951	6,634	0.81	0.72	0.65	0	0.0%
Average	979	5,470	0.84	0.63	0.56	8,445	4.2%
LogMapLite	18	1,822	0.85	0.34	0.21	4,389	2.2%
Tool1	2,140	1,038	0.87	0.23	0.13	649	0.3%

*Uses background knowledge based on the UMLS Metathesaurus which is the basis of the large-bio reference alignments.

Overall, the quality of the results was lower than that observed in the FMA-NCI matching problem, as the matching problem is considerable larger. Like in the FMA-NCI matching problem, the effectiveness off all systems decreases as the ontology size increases from Task 3 to Task 4; and of the systems that completed the former, for example, POMAP was unable to complete the latter.

5.8 Results for the SNOMED-NCI matching problem

Table 10 summarizes the results for the tasks in the SNOMED-NCI matching problem.

AML achieved the best results in terms of both Recall and F-measure in Tasks 5 and 6, while LogMap and AML achieved the best results in terms of precision in Tasks 5 and 6, respectively.

The overall performance of the systems was lower than in the FMA-SNOMED case, as this test case is even larger. Indeed, several systems were unable to complete even the smaller Task 5 within the allotted time: POMAP, SANOM and KEPLER.

As in the previous matching problems, effectiveness decreased as the ontology size increases. Unlike in the FMA-NCI and FMA-SNOMED matching problems, the use of the UMLS Metathesaurus did not positively impact the performance of XMap, which obtained lower results than expected.

Table 10. Results for the SNOMED-NCI matching problem.

System	Time (s)	# Corresp.	Scores			Incoherence	
			Prec.	F-m.	Rec.	Unsat.	Degree
Task 5: small SNOMED and NCI fragments							
AML	669	14,740	0.87	0.80	0.75	≥3,966	≥5.3%
LogMap	207	12,414	0.95	0.80	0.69	≥0	≥0.0%
LogMapBio	2,779	13,205	0.89	0.77	0.68	≥0	≥0.0%
YAM-BIO	2,202	12,959	0.90	0.77	0.68	≥549	≥0.7%
Average	921	12,220	0.89	0.70	0.59	21,264	28.3%
XMap*	106	16,968	0.89	0.69	0.57	≥46,091	≥61.3%
LogMapLite	9	10,942	0.89	0.69	0.57	≥60,450	≥80.4%
Tool1	481	4,312	0.87	0.35	0.22	≥37,797	≥50.2%
Task 6: whole NCI ontology with SNOMED large fragment							
AML	312	13,176	0.90	0.77	0.67	≥720	≥0.4%
YAM-BIO	490	15,027	0.83	0.76	0.70	≥2,212	≥1.2%
LogMapBio	4,728	13,677	0.84	0.73	0.64	≥5	≥0.003%
LogMap	652	12,273	0.87	0.71	0.60	≥3	≥0.002%
LogMapLite	22	12,894	0.80	0.66	0.57	≥150,656	≥79.5%
Average	1,131	13,666	0.84	0.66	0.56	55,496	29.3%
XMap*	563	23,707	0.82	0.66	0.55	≥137,136	≥72.4%
Tool1	1,150	4,911	0.81	0.34	0.22	≥97,743	≥51.6%

*Uses background knowledge based on the UMLS Metathesaurus which is the basis of the large-bio reference alignments.

6 Disease and Phenotype Track (phenotype)

The Pistoia Alliance Ontologies Mapping project team[9] organises this track based on a real use case where it is required to find alignments between disease and phenotype ontologies. Specifically, in the OAEI 2017 edition of this track the selected ontologies are the Human Phenotype Ontology (HPO), the Mammalian Phenotype Ontology (MP), the Human Disease Ontology (DOID), the Orphanet and Rare Diseases Ontology (ORDO), the Medical Subject Headings (MESH) ontology, and the Online Mendelian Inheritance in Man (OMIM) ontology. The extended results for the OAEI 2016 Disease and Phenotype track (previous campaign) are available in [24].

6.1 Test data

The 2017 edition comprises of four tasks requiring the pairwise alignment of:

- Human Phenotype Ontology (HP) to Mammalian Phenotype Ontology (MP);
- Human Disease Ontology (DOID) to the Orphanet Rare Disease Ontology (ORDO);
- Human Phenotype Ontology (HP) to Medical Subject Headings (MESH); and
- Human Phenotype Ontology (HP) to Online Mendelian Inheritance in Man (OMIM).

Currently, mappings between these ontologies are mostly curated by bioinformatics and disease experts who would benefit from automation of their workflows supported by implementation of ontology matching algorithms.

[9] http://www.pistoiaalliance.org/projects/ontologies-mapping/

Table 11. Disease and Phenotype ontology versions and sources

Ontology	Version	Source
HP	2017-06-30	OBO Foundry
MP	2017-06-29	OBO Foundry
DOID	2017-06-13	OBO Foundry
ORDO	v2.4	ORPHADATA
MESH	Hoehndorf's version (2014)	BioPortal
OMIM	UMLS 2016AB	BioPortal

Table 11 summarizes the ontology versions and sources of the ontologies used in the OAEI 2017. Note that the version and source of HP, MP, DOID and ORDO are different from the ones used in 2016.

We have extracted "baseline" reference alignments based on the available BioPortal mappings (July 8, 2017). Most of the BioPortal [38] mappings are generated automatically by the LOOM[10] system, which should only be considered as a baseline since it is incomplete or may contain errors.

6.2 Evaluation setting

We have run the evaluation in a Ubuntu Laptop with an Intel Core i7-4600U CPU @ 2.10GHz x 4 and allocating 15Gb of RAM.

In the OAEI 2017 phenotype track 10 out of 21 participating OAEI 2017 systems have been able to cope with at least one of the tasks with 4 hours.

6.3 Evaluation criteria

Systems have been evaluated according to the following criteria:

- Precision and recall with respect to a consensus alignment automatically generated by voting based on the outputs of all participating systems (we have used vote=2, vote=3 and vote=4).
- Semantic recall with respect to manually generated mappings for several areas of interest (e.g., carbohydrate, obesity and breast cancer).
- Manual assessment of a subset unique mappings (i.e., mappings that are not suggested by other systems).

We have used the OWL 2 reasoner HermiT to calculate the semantic recall. For example, a positive hit will mean that a mapping in the reference has been (explicitly) included in the output mappings or it can be inferred using reasoning from the input ontologies and the output mappings.[11].

6.4 Use of background knowledge

LogMapBio uses BioPortal as mediating ontology provider, that is, it retrieves from BioPortal the most suitable top-10 ontologies for the matching task.

[10] `https://www.bioontology.org/wiki/index.php/BioPortal_Mappings`

[11] Details about the used notion of semantic precision and recall can be found in [24]

Table 12. Disease and Phenotype task completion.

System	HP-MP	DOID-ORDO	HP-MESH	HP-OMIM
AML	✓	✓	✓	✓
DiSMatch	✓	✓	✓	✓
LogMap	✓	✓	✓	✓
LogMapBio	✓	✓	✓	✓
LogMapLite	✓	✓	✓	*empty*
KEPLER	*time*	✓	*time*	*time*
POMAP	✓	✓	*time*	*time*
Tool1	✓	✓	✓	*empty*
XMap	✓	✓	✓	*empty*
YAM-BIO	✓	✓	✓	*empty*

✓: completed; *empty*: produced empty alignment; *error*: runtime error; *time*: timed out (4 hours).

LogMap uses normalisations and spelling variants from the general (biomedical) purpose UMLS Lexicon (a different resource with respect to the UMLS Metathesaurus).

AML has three sources of background knowledge which can be used as mediators between the input ontologies: the Uber Anatomy Ontology (Uberon), the Human Disease Ontology (DOID) and the Medical Subject Headings (MeSH). Additionally, for the HPO-MP test case, it uses the logical definitions of both ontologies, which define some of their classes as being a combination of an anatomic term (i.e., a class from either FMA or Uberon) with a phenotype modifier term (i.e., a class from the Phenotypic Quality Ontology).

YAM-BIO uses as background knowledge a file containing mappings from the DOID and UBERON ontologies to other ontologies like FMA, NCI or SNOMED CT.

DiSMatch estimates the similarity among concepts through textual semantic relatedness. DiSMatch relies on a corpus of relevant biomedical textual resources.

XMAP uses synonyms provided by the UMLS Metathesaurus.

6.5 Results

AML, DiSMatch, LogMap, and LogMapBio produced the most complete results according to both the automatic and manual evaluation.

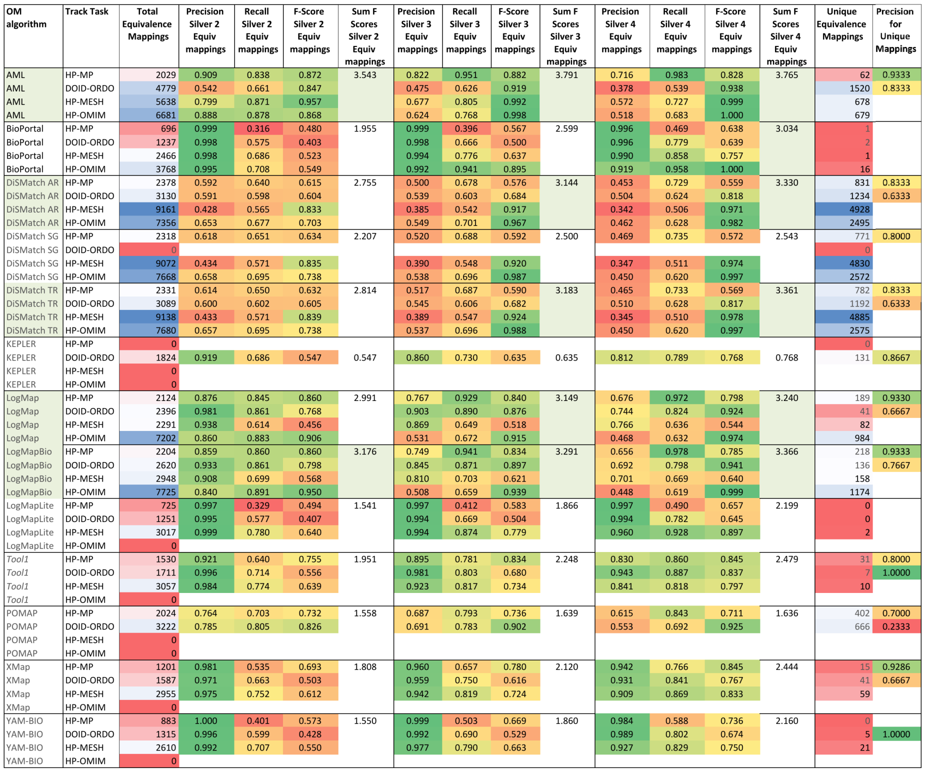
Table 12 summarizes the tasks where each system was able to produce results within a 4-hours time frame.

Results against the consensus alignments Table 13 shows the size of the consensus alignments built with the outputs of the systems participating in the OAEI 2017 campaign. Note that systems participating with different variants only contributed once in the voting, that is, the voting was done by family of systems/variants rather than by individual systems.

Table 3 shows the results achieved by each of the participating systems. We deliberately did not rank the systems since the consensus alignments only allow us to assess how systems perform in comparison with one another. On the one hand, some of the mappings in the consensus alignment may be erroneous (false positives), as all it takes for that is that 2, 3 or 4 systems agree on part of the erroneous mappings they find. On the other hand, the consensus alignments are not complete, as there will likely be correct mappings that no system is able to find, and as we will show in the manual evaluation, there are a number of mappings found by only one system

Table 13. Size of consensus alignments

Task	Vote 2	Vote 3	Vote 4
HP-MP	3,130	2,153	1,780
DOID-ORDO	3,354	2,645	2,188
HP-MESH	4,711	3,847	3,227
HP-OMIM	6,834	4,177	3,462



OM algorithm	Track Task	Total Equivalence Mappings	Precision Silver 2 Equiv mappings	Recall Silver 2 Equiv mappings	F-Score Silver 2 Equiv mappings	Sum F Scores Silver 2 Equiv mappings	Precision Silver 3 Equiv mappings	Recall Silver 3 Equiv mappings	F-Score Silver 3 Equiv mappings	Sum F Scores Silver 3 Equiv mappings	Precision Silver 4 Equiv mappings	Recall Silver 4 Equiv mappings	F-Score Silver 4 Equiv mappings	Sum F Scores Silver 4 Equiv mappings	Unique Equivalence Mappings	Precision for Unique Mappings
AML	HP-MP	2029	0.909	0.838	0.872	3.543	0.822	0.951	0.882	3.791	0.716	0.983	0.828	3.765	62	0.9333
AML	DOID-ORDO	4779	0.542	0.661	0.847		0.475	0.626	0.919		0.378	0.539	0.938		1520	0.8333
AML	HP-MESH	5638	0.799	0.871	0.957		0.677	0.805	0.992		0.572	0.727	0.999		678	
AML	HP-OMIM	6681	0.888	0.878	0.868		0.624	0.768	0.998		0.518	0.683	1.000		679	
BioPortal	HP-MP	696	0.999	0.316	0.480	1.955	0.999	0.396	0.567	2.599	0.996	0.469	0.638	3.034	1	
BioPortal	DOID-ORDO	1237	0.998	0.575	0.403		0.998	0.666	0.500		0.996	0.779	0.639		2	
BioPortal	HP-MESH	2466	0.998	0.686	0.523		0.994	0.776	0.637		0.990	0.858	0.757		1	
BioPortal	HP-OMIM	3768	0.995	0.708	0.549		0.992	0.941	0.895		0.919	0.958	1.000		16	
DiSMatch AR	HP-MP	2378	0.592	0.640	0.615	2.755	0.500	0.678	0.576	3.144	0.453	0.729	0.559	3.330	831	0.8333
DiSMatch AR	DOID-ORDO	3130	0.591	0.598	0.604		0.539	0.603	0.684		0.504	0.624	0.818		1234	0.6333
DiSMatch AR	HP-MESH	9161	0.428	0.565	0.833		0.385	0.542	0.917		0.342	0.506	0.971		4928	
DiSMatch AR	HP-OMIM	7356	0.653	0.677	0.703		0.549	0.701	0.967		0.462	0.628	0.982		2495	
DiSMatch SG	HP-MP	2318	0.618	0.651	0.634	2.207	0.520	0.688	0.592	2.500	0.469	0.735	0.572	2.543	771	0.8000
DiSMatch SG	DOID-ORDO	0													0	
DiSMatch SG	HP-MESH	9072	0.434	0.571	0.835		0.390	0.548	0.920		0.347	0.511	0.974		4830	
DiSMatch SG	HP-OMIM	7668	0.658	0.695	0.738		0.538	0.696	0.987		0.450	0.620	0.997		2572	
DiSMatch TR	HP-MP	2331	0.614	0.650	0.632	2.814	0.517	0.687	0.590	3.183	0.465	0.733	0.569	3.361	782	0.8333
DiSMatch TR	DOID-ORDO	3089	0.600	0.602	0.605		0.545	0.606	0.682		0.510	0.628	0.817		1192	0.6333
DiSMatch TR	HP-MESH	9138	0.433	0.571	0.839		0.389	0.547	0.924		0.345	0.510	0.978		4885	
DiSMatch TR	HP-OMIM	7680	0.657	0.695	0.738		0.537	0.696	0.988		0.450	0.620	0.997		2575	
KEPLER	HP-MP	0													0	
KEPLER	DOID-ORDO	1824	0.919	0.686	0.547	0.547	0.860	0.730	0.635	0.635	0.812	0.789	0.768	0.768	131	0.8667
KEPLER	HP-MESH	0														
KEPLER	HP-OMIM	0														
LogMap	HP-MP	2124	0.876	0.845	0.860	2.991	0.767	0.929	0.840	3.149	0.676	0.972	0.798	3.240	189	0.9330
LogMap	DOID-ORDO	2396	0.981	0.861	0.768		0.903	0.890	0.876		0.744	0.824	0.924		41	0.6667
LogMap	HP-MESH	2291	0.938	0.614	0.456		0.869	0.649	0.518		0.766	0.636	0.544		82	
LogMap	HP-OMIM	7202	0.860	0.883	0.906		0.531	0.672	0.915		0.468	0.632	0.974		984	
LogMapBio	HP-MP	2204	0.859	0.860	0.860	3.176	0.749	0.941	0.834	3.291	0.656	0.978	0.785	3.366	218	0.9333
LogMapBio	DOID-ORDO	2620	0.933	0.861	0.798		0.845	0.871	0.897		0.692	0.798	0.941		136	0.7667
LogMapBio	HP-MESH	2948	0.908	0.699	0.568		0.810	0.703	0.621		0.701	0.669	0.640		158	
LogMapBio	HP-OMIM	7725	0.840	0.891	0.950		0.508	0.659	0.939		0.448	0.619	0.999		1174	
LogMapLite	HP-MP	725	0.997	0.329	0.494	1.541	0.997	0.412	0.583	1.866	0.997	0.490	0.657	2.199	0	
LogMapLite	DOID-ORDO	1251	0.995	0.577	0.407		0.994	0.669	0.504		0.994	0.782	0.645		0	
LogMapLite	HP-MESH	3017	0.999	0.780	0.640		0.994	0.874	0.779		0.960	0.928	0.897		2	
LogMapLite	HP-OMIM	0														
Tool1	HP-MP	1530	0.921	0.640	0.755	1.951	0.895	0.781	0.834	2.248	0.830	0.860	0.845	2.479	31	0.8000
Tool1	DOID-ORDO	1711	0.996	0.714	0.556		0.981	0.803	0.680		0.943	0.887	0.837		7	1.0000
Tool1	HP-MESH	3057	0.984	0.774	0.639		0.923	0.817	0.734		0.841	0.818	0.797		10	
Tool1	HP-OMIM	0														
POMAP	HP-MP	2024	0.764	0.703	0.732	1.558	0.687	0.793	0.736	1.639	0.615	0.843	0.711	1.636	402	0.7000
POMAP	DOID-ORDO	3222	0.785	0.805	0.826		0.691	0.783	0.902		0.553	0.692	0.925		666	0.2333
POMAP	HP-MESH	0														
POMAP	HP-OMIM	0														
XMap	HP-MP	1201	0.981	0.535	0.693	1.808	0.960	0.657	0.780	2.120	0.942	0.766	0.845	2.444	15	0.9286
XMap	DOID-ORDO	1587	0.971	0.663	0.503		0.959	0.750	0.616		0.931	0.841	0.767		41	0.6667
XMap	HP-MESH	2955	0.975	0.752	0.612		0.942	0.819	0.724		0.909	0.869	0.833		59	
XMap	HP-OMIM	0														
YAM-BIO	HP-MP	883	1.000	0.401	0.573	1.550	0.999	0.503	0.669	1.860	0.984	0.588	0.736	2.160	0	
YAM-BIO	DOID-ORDO	1315	0.996	0.599	0.428		0.992	0.690	0.529		0.989	0.802	0.674		5	1.0000
YAM-BIO	HP-MESH	2610	0.992	0.707	0.550		0.977	0.790	0.663		0.927	0.829	0.750		21	
YAM-BIO	HP-OMIM	0														

Fig. 3. Results against consensus alignments with vote 2, 3 and 4.

(and therefore not in the consensus alignments) which are correct. Nevertheless, the results with respect to the consensus alignments do provide some insights into the performance of the systems, which is why we highlighted in the table the 4 systems that produce results closest to the silver standards: AML, DiSMatch, LogMap, and LogMapBio.

Results against manually created mappings The manually generated mappings for six areas (carbohydrate, obesity and breast cancer, urinary incontinence, abnormal heart and Charcot-Marie Tooth disease) include 86 mappings between HP and MP and 175 mappings between DOID and ORDO. Most of them represent subsumption relationships. Tables 14 and 15 shows the results in terms of recall and semantic recall for each of the system. LogMapBio and LogMap

Table 14. Results against manually created mappings: HP-MP task

System	Standard Recall	Semantic Recall
BioPortal (baseline)	*0.20*	*0.51*
AML	0.40	0.62
DiSMatch-ar	0.42	0.65
LogMap	0.38	**0.67**
LogMapBio	0.38	**0.67**
LogMapLt	0.20	0.51
Tool1	0.31	0.60
POMap	0.38	0.65
XMap	0.30	0.60
YAM-BIO	0.22	0.51

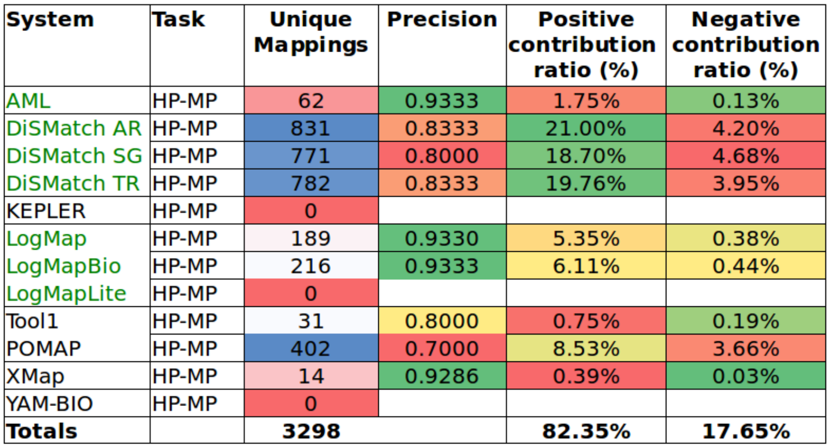
Table 15. Results against manually created mappings: DOID-ORDO task

System	Standard Recall	Semantic Recall
BioPortal (baseline)	*0.13*	*0.14*
AML	0.33	**0.48**
DiSMatch-ar	0.21	0.25
DiSMatch-sg	0.21	0.25
DiSMatch-tr	0.21	0.25
KEPLER	0.13	0.17
LogMap	0.30	0.42
LogMapBio	0.32	0.44
LogMapLt	0.13	0.14
Tool1	0.27	0.30
POMap	0.27	0.30
XMap	0.13	0.14
YAM-BIO	0.13	0.14

obtain the best results in terms of semantic recall in the HP-MP task, while AML obtains the best results in the DOID-ORDO task. The results in both tasks are far from optimal since a large fragment of the manually created mappings have not been (explicitly) identified by the systems nor can be derived via reasoning.

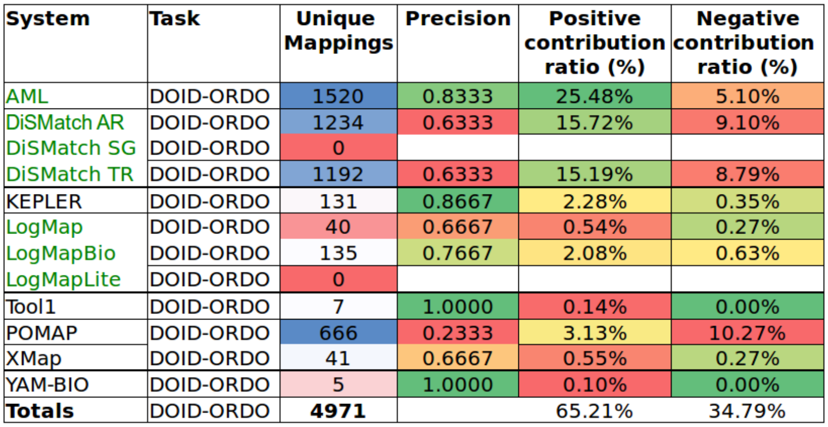
Manual assessment of unique mappings Figures 4 and 5 show the results of the manual assessment to estimate the precision of the unique mappings generated by the participating systems. Unique mappings are correspondences that no other system (explicitly) provided in the output. We manually evaluated up to 30 mappings and we focused the assessment on unique equivalence mappings.

For example LogMap's output contains 189 unique mappings in the HP-MP task. The manual assessment revealed an (estimated) precision of 0.9333. In order to also take into account the number of unique mappings that a system is able to discover, Tables 4 and 5 also include the estimation of the positive and negative contribution of the unique mappings with respect to the total unique mappings discovered by all participating systems.



System	Task	Unique Mappings	Precision	Positive contribution ratio (%)	Negative contribution ratio (%)
AML	HP-MP	62	0.9333	1.75%	0.13%
DiSMatch AR	HP-MP	831	0.8333	21.00%	4.20%
DiSMatch SG	HP-MP	771	0.8000	18.70%	4.68%
DiSMatch TR	HP-MP	782	0.8333	19.76%	3.95%
KEPLER	HP-MP	0			
LogMap	HP-MP	189	0.9330	5.35%	0.38%
LogMapBio	HP-MP	216	0.9333	6.11%	0.44%
LogMapLite	HP-MP	0			
Tool1	HP-MP	31	0.8000	0.75%	0.19%
POMAP	HP-MP	402	0.7000	8.53%	3.66%
XMap	HP-MP	14	0.9286	0.39%	0.03%
YAM-BIO	HP-MP	0			
Totals		**3298**		**82.35%**	**17.65%**

Fig. 4. Unique mappings in the HP-MP task.



System	Task	Unique Mappings	Precision	Positive contribution ratio (%)	Negative contribution ratio (%)
AML	DOID-ORDO	1520	0.8333	25.48%	5.10%
DiSMatch AR	DOID-ORDO	1234	0.6333	15.72%	9.10%
DiSMatch SG	DOID-ORDO	0			
DiSMatch TR	DOID-ORDO	1192	0.6333	15.19%	8.79%
KEPLER	DOID-ORDO	131	0.8667	2.28%	0.35%
LogMap	DOID-ORDO	40	0.6667	0.54%	0.27%
LogMapBio	DOID-ORDO	135	0.7667	2.08%	0.63%
LogMapLite	DOID-ORDO	0			
Tool1	DOID-ORDO	7	1.0000	0.14%	0.00%
POMAP	DOID-ORDO	666	0.2333	3.13%	10.27%
XMap	DOID-ORDO	41	0.6667	0.55%	0.27%
YAM-BIO	DOID-ORDO	5	1.0000	0.10%	0.00%
Totals	DOID-ORDO	**4971**		65.21%	34.79%

Fig. 5. Unique mappings in the DOID-ORDO task.

7 MultiFarm

The MultiFarm data set [33] aims at evaluating the ability of matching systems to deal with ontologies in different natural languages. This data set results from the translation of 7 ontologies from the conference track (cmt, conference, confOf, iasted, sigkdd, ekaw and edas) into 10 languages: Arabic, Chinese, Czech, Dutch, French, German, Italian, Portuguese, Russian, and Spanish. It is composed of 55 pairs of languages (see [33] for details on how the original MultiFarm data set has been generated). For each pair, taking into account the alignment direction (cmt$_{en}$ →confOf$_{de}$ and cmt$_{de}$ →confOf$_{en}$, for instance, as distinct matching tasks), we have 49 matching tasks. The whole data set is composed of 55×49 matching tasks.

7.1 Experimental setting

Part of the data set is used for blind evaluation. This subset includes all matching tasks involving the *edas* and *ekaw* ontologies (resulting in 55×24 matching tasks). As last year, the results reported here are based on the blind data set. Participants were able to test their systems on the available subset of matching tasks (*open evaluation*), available via the SEALS repository. The open subset covers 45×25 tasks. The open subset does not include Italian translations.

We distinguish two types of matching tasks: i) those tasks where two different ontologies (cmt→confOf, for instance) have been translated into two different languages; and ii) those tasks where the same ontology (cmt→cmt) has been translated into two different languages. For the tasks of type ii), good results are not directly related to the use of specific techniques for dealing with cross-lingual ontologies, but on the ability to exploit the identical structure of the ontologies.

This year, 8 systems (out of 22) have participated in the MultiFarm track (i.e., those that have been assigned to the task in the registration phase) : AML, CroLOM, KEPLER, LogMap, LogMapLite, SANOM, WikiV3, and XMAP. LogMapLite does not implement any specific cross-lingual strategy. The number of participants is stable with respect to the last campaign (7 in 2016, 5 in 2015, 3 in 2014, 7 in 2013, and 7 in 2012). For sake of simplicity, we refer in the following to cross-lingual systems those implementing cross-lingual matching strategies and non-cross-lingual systems those without that feature. The reader can refer to the OAEI papers for a detailed description of the strategies adopted by each system. In fact, most of them still adopts a translation step before the matching itself.

For this track, the general comments with respect to the running are : i) CroLOM participated with the same version than last year; ii) LogMap had encountered problems for accessing the Google translator server; iii) KEPLER generated some parsing errors for some pairs; iv) some systems (AML, LogMap and LogMapLite) have generated correspondences with confidence higher than 1.0 (no post-processing has been done in these cases).

7.2 Execution setting and runtime

The systems have been executed on a Windows machine configured with 8GB of RAM running under a i7-7500U CPU 2.70GHz x4 processors. All measurements are based on a single run. As Table 16 shows, we can observe large differences in the time required for a system to complete the 55 x 24 matching tasks. Note as well that the concurrent access to the SEALS repositories during the evaluation period may have an impact in the time required for completing the task.

7.3 Evaluation results

Table 16 presents the aggregated results for the 55×24 matching tasks. They have been computed using the Alignment API 4.6 and can slightly differ from those computed with the SEALS client. We haven't applied any threshold on the results. They are measured in terms of classical precision and recall.

Overall, as expected, systems implementing cross-lingual techniques outperform the non-cross-lingual systems. However, as stated above, this year we did not run all systems and focus on the systems that have been registered for the task. In this task, AML outperforms all other systems in terms of F-measure for task i), keeping its top place in this task. AML is followed by LogMap, CroLOM, KEPLER and WikiV3. With respect to the task ii), AML has relatively low performance, due mainly to some errors in parsing the alignments for which a confidence higher than 1 was generated. KEPLER has provided the higher F-measure for task ii), followed by LogMap, CroLOM and AML. We observe that WikiV3 is able to maintain its performance in both tasks.

Table 16. MultiFarm aggregated results per matcher, for each type of matching task – different ontologies (i) and same ontologies (ii).

System	Time	#pairs	Type (i) – 22 tests per pair				Type (ii) – 2 tests per pair			
			Size	Prec.	F-m.	Rec.	Size	Prec.	F-m.	Rec.
AML	677	55	8.21	.72(.72)	.46(.46)	.35(.35)	45.54	.89(.96)	.26(.28)	.16(.17)
CroLOM	5501	55	8.56	.55(.55)	.36(.36)	.28(.28)	38.76	.89(.90)	.40(.40)	.26(.27)
KEPLER	2180	55	10.63	.43(.43)	.31(.31)	.25(.25)	58.34	.90(.90)	.52(.52)	.38(.38)
LogMap	57	55	6.99	.73(.73)	.37(.37)	.25(.25)	46.80	.95(.96)	.42(.43)	.28(.28)
LogMapLite	38	55	1.16	.36(.36)	.04(.04)	.02(.02)	94.5	.02(.02)	.01(.03)	.01(.02)
SANOM	22	30	2.86	.43(.79)	.13(.25)	.08(.15)	8.33	.54(.99)	.06(.12)	.03(.06)
WikiV3	1343	55	11.89	.30(.30)	.25(.25)	.21(.21)	29.37	.62(.62)	.23(.23)	.14(.14)
XMAP	102	27	3.84	.24(.50)	.06(.14)	.04(.09)	15.76	.66(.91)	.10(.14)	.06(.09)

Time is measured in minutes (for completing the 55×24 matching tasks); #pairs indicates the number of pairs of languages for which the tool is able to generated (non empty) alignments; size indicates the average of the number of generated correspondences for the tests where an (non empty) alignment has been generated. Two kinds of results are reported: those do not distinguishing empty and erroneous (or not generated) alignments and those—indicated between parenthesis—considering only non empty generated alignments for a pair of languages.

With respect to the pairs of languages for test cases of type i), for the sake of brevity, we do not present the results for the 55 pairs. The reader can refer to the OAEI results web page for the detailed results. 5 cross-lingual systems out of 7 were able to deal with all pairs of languages (AML, CroLOM, KEPLER, LogMap and WikiV3). While the only non-specific system was able to generate non empty (but erroneous) results for all pairs, specific systems as SANOM and XMap have problems to deal with ar, cn and ru languages and hence were not able to generate alignments for most pairs involving these languages. This behaviour has also been observed in the last campaign for specific systems.

For the group of systems implementing cross-lingual strategies, their top F-measure include the pairs es-it (AML), nl-pt (CroLOM), de-pt (KEPLER), en-nl (LogMap), es-it (SANOM), it-pt (WikiV3), es-pt (XMap). We can observe that most of the systems better deal with the pairs involving pt, it, es, nl, de and en languages. This may due to the coverage or performance of the resources and translations for these languages, together with the fact that dealing with comparable languages[12] can make the task easier. In fact, we can also observe that for most systems, the worst results have been produced for the pairs involving ar, cn, cz and ru. The exceptions are SANOM and XMap, for which, worst results also include the pairs es, nl and pt or fr, en and it, respectively.

With respect to the only non cross-lingual systems, LogMapLite, it in fact takes advantage of comparable languages, in the absence of specific strategies. This can be corroborated by the fact that it has generated its best F-measure for the pairs de-en, es-pt, it-pt, es-it. This (expected) fact has been observed along the campaigns.

[12] An example of comparable natural languages is English and German, both belonging to the Germanic language family. Comparable natural languages can also be languages that are not from the same language family. For example, Italian belonging to the Romance language family, and German belonging to the Germanic language family can still be compared using string comparison techniques such as edit distance, as they are both alphabetic letter-based with comparable graphemes. An example of natural languages that are not comparable in this context can be Chinese and English, where the former is logogram-based and the latter is alphabetic letter-based [12]

Comparison with previous campaigns. The number of participants implementing cross-lingual strategies remains stable this year with respect to the last campaigns (7 in 2016, 5 in 2015, 3 in 2014, 7 in 2013 and 2012 and 3 in 2011). 4 systems have also participated last year (AML, LogMap, CroLOM, and XMap) and we count 3 new systems (KEPLER, SANOM, and WikiV3). Comparing the results from last year, in terms F-measure and with respect to the blind evaluation (cases of type i), AML maintains its performance, with a very little increase (.46 in 2017, .45 in 2016 and .47 in 2015). CroLOM, LogMap, and XMAP maintained their performance (.36, .37 and .06, respectively). The newcomer WikiV3 obtained stable results for both kinds of tasks, but with a F-measure below AML, LogMap, CroLOM and KEPLER. For the task ii), we can observe that KEPLER (.52) outperforms LogMap (.44), the best system from last year, in terms of F-measure for this task.

7.4 Conclusion

From 22 participants, 8 were evaluated in MultiFarm. In terms of performance, the F-measure for blind tests remains relatively stable across campaigns. AML and LogMap keep their positions with respect to the previous campaigns, followed by the CroLOM and KEPLER. Still, all systems privilege precision in detriment to recall and the results are below the ones obtained for the Conference original dataset. We can observe as well that the systems are not able to provide good results or deal with pairs involving specific languages, as ar, cn and ru. As last years, still cross-lingual approaches are mainly based on translation strategies and the combination of other resources (like cross-lingual links in Wikipedia, BabelNet, etc.) and strategies (machine learning, indirect alignment composition) remains underexploited. As last year, the evaluation has been conducted only on the blind set (results have not been reported for the open data set). As future work, we plan to compare the performance of the systems on both multilingual and cross-lingual settings.

8 Interactive matching

The interactive matching track was organized at OAEI 2017 for the fifth time. The goal of this evaluation is to simulate interactive matching [36, 14], where a human expert is involved to validate correspondences found by the matching system. In the evaluation, we look at how interacting with the user improves the matching results. Currently, this track does not evaluate the user experience or the user interfaces of the systems.

8.1 Datasets

The Interactive track uses four OAEI datasets: Anatomy (Section 3), Conference (Section 4), LargeBio (Section 5), and Phenotype (Section 6). For details on the datasets, please refer to their respective sections.

8.2 Experimental setting

The Interactive track relies on the SEALS client's *Oracle* class to simulate user interactions. An interactive matching system can present a correspondence to the oracle, which will tell the system whether that correspondence is right or wrong. This year we have extended this functionality by allowing a user to present a collection of mappings simultaneously to the oracle. If a system presents up to three mappings together and each mapping presented has a mapped entity (i.e., class or property) in common with at least one other mapping presented, the oracle counts this as a single interaction, under the rationale that this corresponds to a scenario where a user is asked to choose between conflicting candidate mappings.

To simulate the possibility of user errors, the oracle can be set to reply with a given error probability (randomly, from a uniform distribution). We evaluated systems with four different error rates: 0.0 (perfect user), 0.1, 0.2, and 0.3.

The evaluations of the Conference and Anatomy datasets were run on a server with 3.46 GHz (6 cores) and 8GB RAM allocated to the matching systems. Each system was run ten times and the final result of a system for each error rate represents the average of these runs. For the Conference dataset with the ra1 alignment, precision and recall correspond to the micro-average over all ontology pairs, whereas the number of interactions represent the total number of interactions for all the pairs. Both are averaged for the ten runs.

The Phenotype and Largebio evaluation was run on a Ubuntu Laptop with an Intel Core i7-4600U CPU @ 2.10GHz x 4 and allocating 15Gb of RAM. Each system was run only one time due to the time required to run some of the systems. Since errors are randomly introduced we expect minor variations between runs. Nevertheless, the Phenotype and Largebio tasks involve large ontologies and a comparatively large number of questions, hence the variations between runs are expected to be mostly negligible.

8.3 Evaluation

For the sake of brevity, we present only the results for the Anatomy, Conference, and LargeBio tasks. For the Phenotype tasks, please refer to the OAEI website [13]. Table 17 and Figure 6 show the results for the Anatomy and Conference datasets, and Table 18 and Figure 7 show the results for the LargeBio tasks.

The tables include the following information (column names within parentheses):

- The number of unsatisfiable classes resulting from the alignments computed as detailed in Section 5 - only for the LargeBio data set.
- The performance of the system: Precision (Prec.), Recall (Rec.) and F-measure (F-m.) with respect to the fixed reference alignment, as well as Recall+ (Rec.+) for the Anatomy task (as detailed in Section 3). To facilitate the assessment of the impact of user interactions, we also provide the performance results from the original tracks, without interaction (line with Error NI).
- To ascertain the impact of the oracle errors, we provide the performance of the system with respect to the oracle (i.e., the reference alignment as modified by the errors introduced by the oracle: Precision oracle (Prec. oracle), Recall oracle (Rec. oracle) and F-measure oracle (F-m. oracle). For a perfect oracle these values match the actual performance of the system.
- Total requests (Tot Reqs.) represents the number of distinct user interactions with the tool, where each interaction can contain one to three conflicting mappings, that could be analysed simultaneously by a user.
- Distinct mappings (Dist. Mapps) counts the total number of mappings for which the oracle gave feedback to the user (regardless of whether they were submitted simultaneously, or separately).
- Finally, the performance of the oracle itself with respect to the errors it introduced can be gauged through the positive precision (Pos. Prec.) and negative precision (Neg. Prec.), which measure respectively the fraction of positive and negative answers given by the oracle that are correct. For a perfect oracle these values are equal to 1 (or 0, if no questions were asked).

[13] http://oaei.ontologymatching.org/2017/results/interactive/

The figures show the time intervals between the questions to the user/oracle for the different systems and error rates. Different runs are depicted with different colours.

8.4 Discussion

The matching systems that participated in this track employ different user-interaction strategies. While LogMap, XMap and AML make use of user interactions exclusively in the post-matching steps to filter their candidate mappings, ALIN can also add new candidate mappings to its initial set. LogMap and AML both request feedback on only selected mapping candidates (based on their similarity patterns or their involvement in unsatisfiabilities) and AML presents one mapping at a time to the user. XMap also presents one mapping at a time and asks mainly about false mappings. ALIN and LogMap can both ask the oracle to analyse several conflicting mappings simultaneously.

The performance of the systems usually improves when interacting with a perfect oracle in comparison with no interaction. The one exception is XMap in the Conference dataset, because it is barely interactive in this dataset. In general, XMap performs very few requests to the oracle compared to the other systems, except in the SNOMED-NCI task, where it makes the most requests. Thus, it is also the system that improves the least with user interaction. On the other end of the spectrum, ALIN is the system that improves the most, not only because it makes a high number of oracle requests (the most in Anatomy and Conference) but also because its non-interactive performance was the lowest of the interactive systems, and thus the easiest to improve.

Although systems' performance deteriorates when the error rate increases, there are still benefits from the user interaction—some of the systems' measures stay above their non-interactive values even for the larger error rates. Naturally, the more a system relies on the oracle, the more its performance tends to be affected by its errors.

The impact of the oracle's errors is linear for ALIN, AML and for XMap in most tasks, as the F-measure according to the oracle remains approximately constant across all error rates. It is supra-linear for LogMap in all data sets, and for XMap in the SNOMED-NCI task, as the F-measure according to the oracle decreases as the error rate increases. This means that the latter systems are deliberately or implicitly letting the oracle's replies affect their selection of mappings beyond those they asked about, and thus propagating the oracle's errors.

Two models for system *response times* are frequently used in the literature [10]: Shneiderman and Seow take different approaches to categorise the response times. Shneiderman takes a task-centred view and sorts the response times in four categories according to task complexity: typing, mouse movement (50-150 ms), simple frequent tasks (1 s), common tasks (2-4 s) and complex tasks (8-12 s). He suggests that the user is more tolerable to delays with the growing complexity of the task at hand. Unfortunately, no clear definition is given for how to define the task complexity. Seow's model looks at the problem from a user-centred perspective by considering the user expectations towards the execution of a task: instantaneous (100-200 ms), immediate (0.5-1 s), continuous (2-5 s), captive (7-10 s). Ontology alignment is a cognitively demanding task and can fall into the third or fourth categories in both models. In this regard the response times (request intervals as we call them above) observed in all data sets fall into the tolerable and acceptable response times, and even into the first categories, in both models. The request intervals for AML, LogMap and XMAP stay at a few milliseconds for most data sets. ALIN's request intervals are higher, but still in the tenth of second range. It could be the case, however, that a user would not be able to take advantage of these low response times because the task complexity may result in higher user response time (i.e., the time the user needs to respond to the system after the system is ready).

Regarding the number of unsatisfiable classes resulting from the alignments we observe some expected variations as the error increases. We note that, with interaction, the alignments

Table 17. Interactive matching results for the Anatomy and Conference datasets

Tool	Error	Prec.	Rec.	F-m.	Rec.+	Prec. oracle	Rec. oracle	F-m. oracle	Tot. Reqs.	Dist. Mapps	Pos. Prec.	Neg. Prec.
Anatomy Dataset												
ALIN	NI	0.985	0.339	0.504	0.0	–	–	–	–	–	–	–
	0.0	0.993	0.794	0.882	0.454	0.993	0.794	0.882	939	1472	1.0	1.0
	0.1	0.94	0.745	0.831	0.403	0.993	0.79	0.88	905	1352	0.905	0.8977
	0.2	0.895	0.703	0.787	0.358	0.993	0.788	0.879	891	1311	0.824	0.796
	0.3	0.846	0.649	0.735	0.301	0.993	0.781	0.874	882	1266	0.734	0.668
AML	NI	0.95	0.936	0.943	0.832	–	–	–	–	–	–	–
	0.0	0.968	0.948	0.958	0.862	0.968	0.948	0.958	241	240	1.0	1.0
	0.1	0.956	0.946	0.95	0.856	0.969	0.949	0.959	266	264	0.73	0.972
	0.2	0.939	0.942	0.94	0.849	0.969	0.951	0.96	283	280	0.513	0.93
	0.3	0.922	0.939	0.931	0.843	0.97	0.952	0.961	310	308	0.359	0.902
LogMap	NI	0.911	0.846	0.877	0.593	–	–	–	–	–	–	–
	0.0	0.982	0.846	0.909	0.595	0.982	0.846	0.909	388	1164	1.0	1.0
	0.1	0.962	0.83	0.891	0.564	0.966	0.803	0.877	388	1164	0.748	0.964
	0.2	0.944	0.823	0.88	0.552	0.945	0.762	0.843	388	1164	0.566	0.927
	0.3	0.931	0.82	0.872	0.544	0.92	0.722	0.809	388	1164	0.431	0.879
XMap	NI	0.926	0.863	0.893	0.639	–	–	–	–	–	–	–
	0.0	0.927	0.865	0.895	0.644	0.927	0.865	0.895	35	35	1.0	1.0
	0.1	0.927	0.865	0.895	0.644	0.927	0.863	0.894	35	35	0.602	0.964
	0.2	0.927	0.865	0.895	0.644	0.927	0.862	0.893	35	35	0.422	0.964
	0.3	0.927	0.865	0.895	0.644	0.927	0.861	0.893	35	35	0.278	0.93
Conference Dataset												
ALIN	NI	0.892	0.272	0.417	–	–	–	–	–	–	–	–
	0.0	0.957	0.731	0.829	–	0.957	0.731	0.829	329	571	1.0	1.0
	0.1	0.804	0.669	0.73	–	0.961	0.737	0.834	321	549	0.752	0.966
	0.2	0.669	0.622	0.645	–	0.965	0.751	0.845	313	534	0.558	0.93
	0.3	0.577	0.56	0.568	–	0.966	0.752	0.845	302	517	0.431	0.875
AML	NI	0.841	0.659	0.739		–	–	–	–	–	–	–
	0.0	0.912	0.711	0.799	–	0.912	0.711	0.799	271	270	1.0	1.0
	0.1	0.841	0.701	0.765	–	0.923	0.732	0.816	282	275	0.704	0.975
	0.2	0.768	0.672	0.717	–	0.925	0.745	0.825	292	279	0.538	0.92
	0.3	0.713	0.651	0.68	–	0.929	0.751	0.83	291	274	0.45	0.877
LogMap	NI	0.818	0.59	0.686	–	–	–	–	–	–	–	–
	0.0	0.886	0.61	0.723	–	0.886	0.61	0.723	82	246	1.0	1.0
	0.1	0.851	0.598	0.702	–	0.855	0.573	0.686	82	246	0.698	0.978
	0.2	0.821	0.585	0.684	–	0.829	0.542	0.656	82	246	0.507	0.941
	0.3	0.795	0.581	0.671	–	0.807	0.518	0.631	82	246	0.363	0.902
XMap	NI	0.837	0.57	0.678	–	–	–	–	–	–	–	–
	0.0	0.837	0.57	0.678	–	0.837	0.57	0.678	4	4	0.0	1.0
	0.1	0.837	0.57	0.678	–	0.837	0.57	0.678	4	4	0.0	1.0
	0.2	0.837	0.57	0.678	–	0.837	0.569	0.677	4	4	0.0	1.0
	0.3	0.837	0.57	0.678	–	0.837	0.569	0.678	4	4	0.0	1.0

NI stands for non-interactive, and refers to the results obtained by the matching system in the original track.

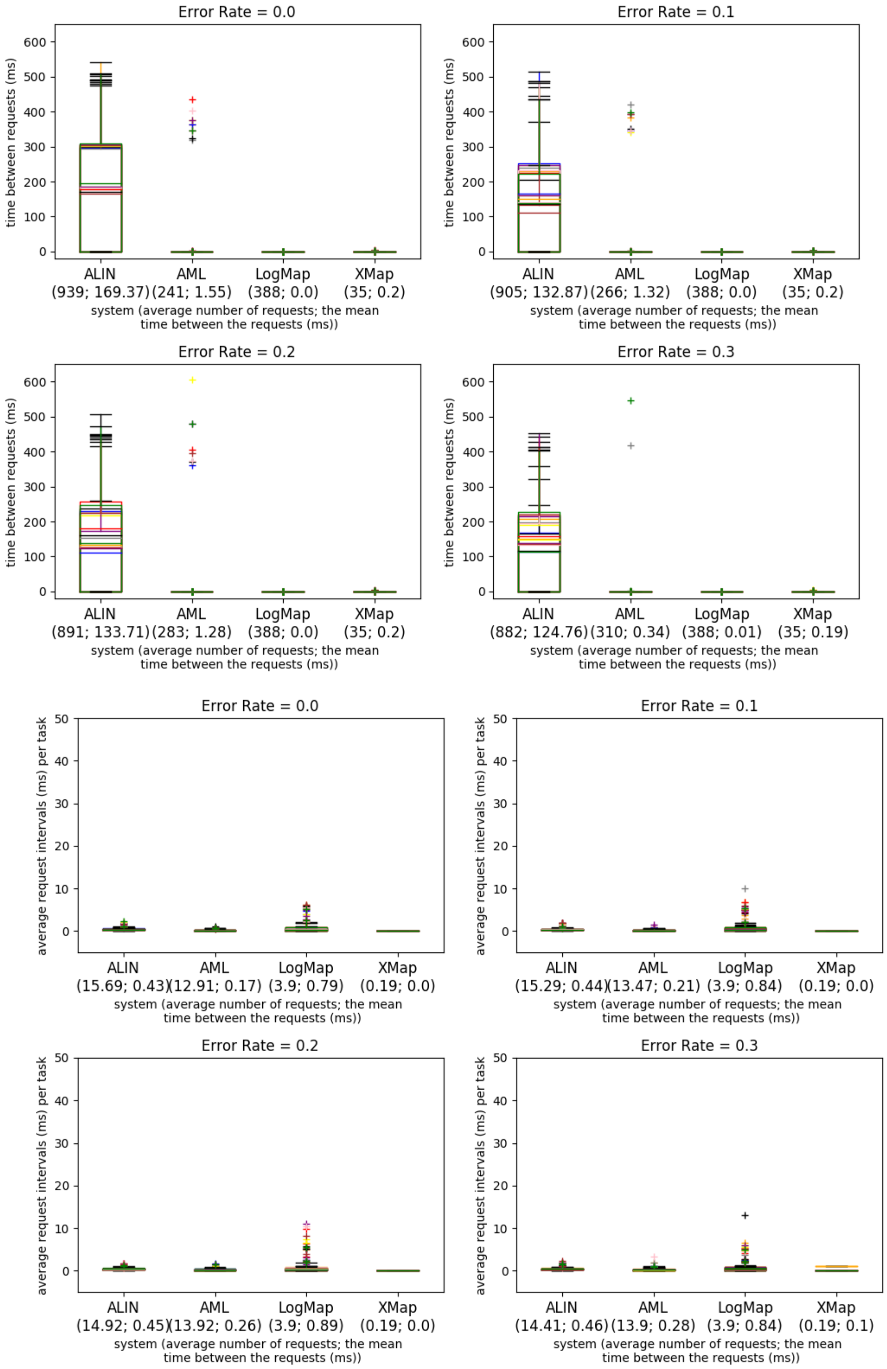


Fig. 6. Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

Table 18. Interactive matching results for the LargeBio dataset

						Prec.	Rec.	F-m.	Tot.	Dist.	Pos.	Neg.
Tool	Error	Unsat.	Prec.	Rec.	F-m.	oracle	oracle	oracle	Reqs.	Mapps	Prec.	Prec.
FMA-NCI Small Dataset												
ALIN	NI	N/A	0.995	0.455	0.624	–	–	–	–	–	–	–
	0.0	2	0.996	0.63	0.772	0.996	0.63	0.772	653	1,019	1	1
	0.1	85	0.971	0.614	0.752	0.996	0.63	0.772	629	932	0.908	0.907
	0.2	152	0.958	0.593	0.733	0.996	0.624	0.767	605	881	0.855	0.788
	0.3	91	0.937	0.58	0.716	0.996	0.623	0.767	589	855	0.772	0.696
AML	NI	2	0.963	0.902	0.932	–	–	–	–	–	–	–
	0.0	2	0.99	0.913	0.95	0.99	0.913	0.95	449	447	1	1
	0.1	222	0.98	0.908	0.943	0.99	0.914	0.95	497	484	0.896	0.936
	0.2	2	0.974	0.894	0.932	0.987	0.91	0.947	450	450	0.794	0.768
	0.3	2	0.966	0.894	0.929	0.981	0.911	0.945	450	450	0.751	0.734
LogMap	NI	2	0.944	0.897	0.92	–	–	–	–	–	–	–
	0.0	2	0.992	0.901	0.944	0.992	0.901	0.944	1,131	1,131	1	1
	0.1	2	0.98	0.881	0.928	0.983	0.892	0.935	1,209	1,209	0.942	0.909
	0.2	2	0.967	0.874	0.918	0.964	0.875	0.917	1,247	1,247	0.837	0.84
	0.3	2	0.963	0.872	0.915	0.935	0.849	0.89	1,327	1,327	0.727	0.776
XMap	NI	2	0.977	0.901	0.937	–	–	–	–	–	–	–
	0.0	2	0.991	0.9	0.943	0.991	0.9	0.943	188	188	1	1
	0.1	2	0.988	0.895	0.939	0.99	0.9	0.943	187	187	0.962	0.819
	0.2	2	0.988	0.892	0.938	0.99	0.899	0.942	187	187	0.939	0.753
	0.3	2	0.985	0.887	0.933	0.99	0.899	0.942	188	188	0.851	0.628
SNOMED-NCI Small Dataset												
AML	NI	3,966	0.904	0.713	0.797	–	–	–	–	–	–	–
	0.0	0	0.972	0.726	0.831	0.972	0.726	0.831	2,730	2,730	1	1
	0.1	0	0.967	0.717	0.823	0.972	0.724	0.83	2,730	2,730	0.942	0.857
	0.2	0	0.961	0.707	0.815	0.972	0.721	0.828	2,730	2,730	0.88	0.732
	0.3	0	0.955	0.697	0.806	0.972	0.719	0.827	2,730	2,730	0.818	0.622
LogMap	NI	0	0.922	0.663	0.771	–	–	–	–	–	–	–
	0.0	0	0.985	0.669	0.797	0.985	0.669	0.797	5,596	5,596	1	1
	0.1	16	0.974	0.651	0.78	0.971	0.656	0.783	6,201	6,201	0.945	0.855
	0.2	16	0.965	0.64	0.77	0.948	0.639	0.763	6,737	6,737	0.859	0.766
	0.3	16	0.959	0.635	0.764	0.92	0.62	0.741	7,159	7,159	0.753	0.693
XMap	NI	46,091	0.911	0.564	0.697	–	–	–	–	–	–	–
	0.0	35,869	0.924	0.59	0.72	0.924	0.59	0.72	11,932	11,689	1	1
	0.1	35,455	0.923	0.591	0.721	0.84	0.568	0.678	11,931	11,694	0.99	0.602
	0.2	35,968	0.921	0.591	0.72	0.754	0.541	0.63	11,911	11,682	0.975	0.41
	0.3	36,619	0.919	0.592	0.72	0.676	0.514	0.584	11,903	11,693	0.953	0.297

NI stands for non-interactive, and refers to the results obtained by the matching system in the original track. ALIN was unable to complete the SNOMED-NCI task.

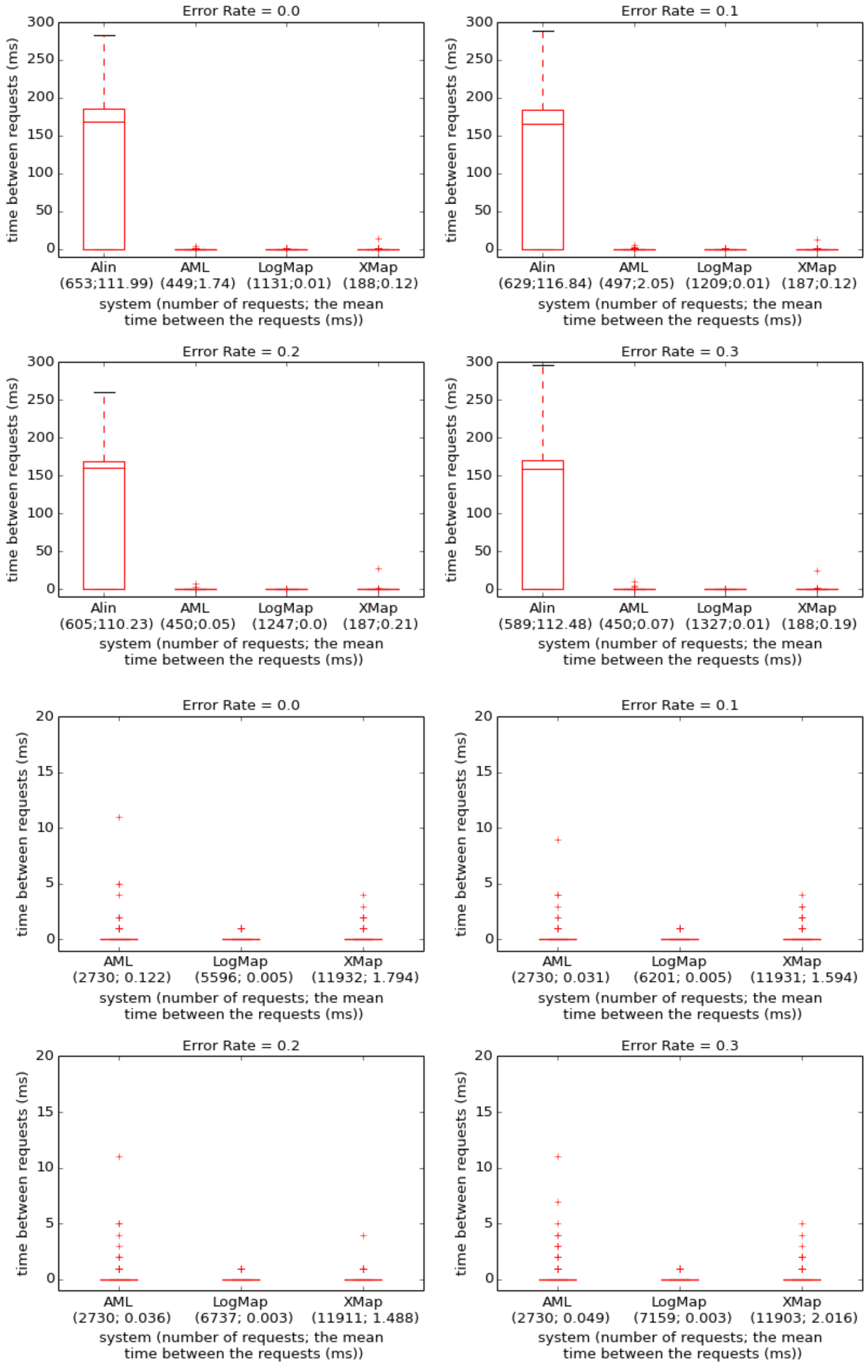


Fig. 7. Time intervals between requests to the user/oracle for the FMA-NCI (top 4 plots) and SNOMED-NCI (bottom 4 plots) datasets from the LargeBio track. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the number of requests and the mean time between the requests.

produced by the systems are typically larger than without interaction, which makes the repair process harder. The introduction of oracle errors complicates the process further, and may make an alignment irreparable if the system follows the oracle's feedback blindly.

9 Instance matching

The instance matching track aims at evaluating the performance of matching tools when the goal is to detect the degree of similarity between pairs of items/instances expressed in the form of OWL Aboxes. The track is organized in two independent tasks called *SYNTHETIC* and *DOREMUS*. Each test is based on two datasets called source and target and the goal is to discover the matching pairs (i.e., mappings) among the instances in the source dataset and the instances in the target dataset.

For the sake of clarity, we split the presentation of he task results in two different subsections.

9.1 SYNTHETIC task

Task data The SYNTHETIC datasets are produced using SPIMBENCH [40] with the aim to generate descriptions of the same entity where *value-based*, *structure-based* and *semantics-aware* transformations are employed on source data in order to create the target data.

The value-based transformations consider mainly typographical errors and different data formats, the structure-based transformations implement transformations applied on the structure of object and datatype properties and the semantics-aware transformations concern the instance level and take into account schema information. The latter are used to examine if the matching systems take into account RDFS and OWL constructs in order to discover correspondences between instances that can be found only by considering schema information.

We stress that an instance in the source dataset can have none or one matching counterpart in the target dataset. A dataset is composed of a Tbox and a corresponding Abox. Source and target datasets share almost the same Tbox (differences are found in the properties due to the employed structure-based transformations). The *Sandbox* scale is 10K triples $\approx$ 380 instances while the *Mainbox* scale is 50K triples $\approx$ 1800 instances. We asked the participants to match the creative works (news items, blogposts and programmes) in the source dataset against the instances of the corresponding class in the target dataset.

Results The participants of the SYNTHETIC task are the AgreementMakerLight (AML), I-Match, Legato and LogMap systems. In order to evaluate those systems we built a ground truth containing the set of expected links where an instance i_1 in the source dataset is associated with an instance j_1 in the target dataset that has been generated as a modified description of i_1. The value-based, structure-based and semantics-aware transformations were applied on different triples of the source dataset pertaining to one class instance.

The systems were judged on the basis of the *precision*, *recall* and *F-measure* results shown in Table 19. LogMap and Legato produce links that are very often correct (resulting in a good precision) but fail to capture a large number of the expected links (resulting in a lower recall). In the case of AML and I-Match systems, the probability of capturing a correct link is high, but the probability of a retrieved link to be correct is lower, resulting in a high (almost perfect) recall but a low precision. Regarding the size of the dataset, LogMap and Legato systems have better results for the Sandbox dataset. On the other hand, AML and I-Match systems exhibit the same performance for both the Sandbox and Mainbox datasets.

Table 19. SYNTHETIC task results

System	Sandbox task			Mainbox task		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
AML	0.849	1.000	0.918	0.855	1.000	0.922
I-Match	0.854	0.997	0.920	0.856	0.997	0.921
Legato	0.980	0.730	0.840	0.970	0.700	0.810
LogMap	0.938	0.763	0.841	0.893	0.709	0.790

9.2 DOREMUS task

Task data The DOREMUS task, having its second appearance at the OAEI, contains real world datasets coming from two major French cultural institutions – The BnF (French National Library) and the PP (Philharmonie de Paris). The data are about classical music works and follow the DOREMUS model (one single vocabulary for both datasets) issued from the DOREMUS project.[14] Each data entry, or instance, is a bibliographical record about a musical piece, containing properties such as the composer, the title(s) of the work, the year of creation, the key, the genre, the instruments, to name a few. These data have been converted to RDF from their original UNI- and INTER-MARC formats and anchored to the DOREMUS ontology and a set of domain controlled vocabularies by the help of the *marc2rdf* converter,[15] developed for this purpose within the DOREMUS Project (for more details on the conversion method and on the ontology we refer to [1] and [31]). Note that these data are highly heterogeneous. We have selected works described both at the BnF and at the PP with different degrees of heterogeneity in their descriptions. The datasets have been selected for the purposes of two sub-tasks.

Heterogeneities (HT): This sub-task consists in aligning two datasets, BnF-1 and PP-1, containing about 238 instances each, by discovering 1:1 equivalence relations between them. There are different types of heterogeneities that these data manifest, identified by music library experts, such as multilingualism, differences in catalogs, differences in spelling, different degrees of description, etc. The goal is to test the ability of linking tools to cope with these heterogeneities. The participants are asked to map only instances of the $F22\_Self - Contained\_Expression$ class.

False Positives Trap (FPT): This sub-task consists in correctly disambiguating the instances contained in two datasets of small sizes (75 instances each), BnF-2 and PP-2, by discovering 1:1 equivalence relations between the instances that they contain. Librarian experts have selected several groups of music works with highly similar descriptions across the two datasets, where there exist only one correct match in each group. The goal is to challenge the linking tools capacity to avoid the generation of false positives and match correctly instances in the presence of highly similar but yet distinct candidates. The participants are asked to map only instances of the $F22\_Self - Contained\_Expression$ class.

Results Five systems participated and returned results on the DOREMUS track: AML, I-Match, *Legato*, LogMap and NjuLink. Two systems stand out, outperforming significantly the other participants on both sub-tasks – Legato and NjuLink, both achieving F-measures of over 0.9

[14] http://www.doremus.org

[15] https://github.com/DOREMUS-ANR/marc2rdf

(NjuLink leading on HT and Legato - on FP-trap). Both tasks appear to be fairly challenging for the majority of the systems, with average F-measures of 0.636 for HT task and 0.565 for the FP-trap task.

Table 20. Results of the DOREMUS task

System	HT task			FP-Trap task		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
AML	0.851	0.479	0.613	0.914	0.427	0.582
I-Match	0.680	0.071	0.129	1.00	0.053	0.101
Legato	0.930	0.920	**0.930**	1.00	0.980	**0.990**
LogMap	0.406	0.882	0.556	0.119	0.880	0.210
NjuLink	0.966	0.945	**0.955**	0.959	0.933	**0.946**

10 HOBBIT Link Discovery

In this track, two benchmark generators are proposed to deal with *link discovery* for spatial data represented as *trajectories* i.e., sequences of longitude, latitude pairs. This new track is using the HOBBIT platform[16] and follows different instructions than the SEALS-based tracks.

We use TomTom[17] datasets in order to create the benchmark. TomTom datasets contain representations of traces (GPS fixes). Each trace consists of a number of points. Each point has a timestamp, longitude, latitude and speed. The points are sorted in ascending order by the timestamp of the corresponding GPS fix. Each task of the HOBBIT Link Discovery Track is composed of two datasets with different number of instances to match, namely the Sandbox and the Mainbox.

The HOBBIT Link Discovery track comprises of two tasks:

- *Task 1 (Linking)* measures how well the systems can match traces that have been modified using string-based approaches along with addition and deletion of intermediate points. Since TomTom datasets only contain coordinates, in order to apply string-based modifications implemented in LANCE [41] we have replaced a number of those points with labels retrieved from Linked Data spatial datasets using the Google Maps[18], Foursquare[19] and Nominatim Openstreetmap[20] APIs. This task also contains modifications on date and coordinate formats. An instance in the source dataset has one matching counterpart in the target dataset. For the *Linking Task*, the Sandbox scale is 100 instances while the Mainbox scale is 5K instances. We asked the participants to match traces in the source and the target datasets.
  The participants of the Linking task are AgreementMakerLight (AML) and OntoIdea systems. For evaluation, we built a ground truth containing the set of expected links where an instance i_1 in the source dataset is associated with an instance j_1 in the target dataset that has been generated as an altered description of i_1.
  The way that the transformations were done, was to apply value-based, and structure-based transformations on different triples pertaining to instances of class Trace.

[16] `https://project-hobbit.eu/outcomes/hobbit-platform/`
[17] `https://www.tomtom.com/`
[18] `https://developers.google.com/maps/`
[19] `https://developer.foursquare.com/`
[20] `http://nominatim.openstreetmap.org/`

Table 21. HOBBIT Link Discovery Linking Task

System	Precision	Recall	F-measure	Run Time
Sandbox task				
AML	1.000	1.000	1.000	11722
OntoIdea	0.990	0.990	0.990	19806
Mainbox task				
AML	1.000	1.000	1.000	134456
OntoIdea	Platform Time Limit (75 mins)			

The systems were judged on the basis of *precision*, *recall*, *F-measure* and *runtime* results that are shown in Table 21. Both AML and OntoIdea systems return high precision and recall capturing all the correct links. Regarding runtime, for the Sandbox dataset, AML needs less time than OntoIdea and for the Mainbox dataset, AML completes the task with perfect results in contrast to OntoIdea that was not able to complete it and stopped when it hit the platform time limit (75 mins). Datasets, reference alignments, and task results are available on the HOBBIT website: `https://project-hobbit.eu/challenges/om2017/`.

– *Task 2 (Spatial)* measures how well the systems can identify the DE-9IM (Dimensionally Extended nine-Intersection Model) topological relations. The supported spatial relations are the following: *Equals, Disjoint, Touches, Contains/Within, Covers/CoveredBy, Intersects, Crosses, Overlaps*. The traces are represented in the Well-known text (WKT) format. For each relation, a different pair of source and target datasets is given to the participants.

Given a LineString source geometry s, a LineString target geometry t and a DE-9IM topological relation r, we ask the participants to match an instance from s with one or more instances in t such as their Intersection Matrix follows the definition of r. For evaluation, we built a ground truth using RADON [42] containing the set of expected links where an instance i_1 in the source dataset is associated with one or more instances in the target dataset that has been generated as an altered description of i_1. For the *Spatial Task*, the Sandbox scale is 10 instances and the Mainbox scale is 2K instances.

The participants to the Spatial task are AgreementMakerLight (AML), OntoIdea, Rapid Discovery of Topological Relations (RADON) and Silk systems.

The systems were judged on the basis of *precision*, *recall*, *F-measure* and *runtime* results shown in Table 22 and Figures 8 and 9. We should mention that we are only presenting the time performance and not precision, recall and f-measure as all were equal to 1.0 except *OntoIdea* that reports for the *Touches* and *Overlaps* relations value 0.99. Moreover, Silk is not participating in relations *Covers* and *Covered By* and OntoIdea is not participating in relation *Disjoint*.

From the results we can observe that:

- **OntoIdea** has the best performance in the Sandbox dataset *but* in the Mainbox dataset the runtime increases and the system seems to not be able to handle large datasets easily.
- **Silk** also seems to have a similar behaviour as **OntoIdea**.
- **RADON** and **AML** systems seem to handle the growth of the dataset size smoother.
- **AML** does not provide any results for the *Disjoint* relation since it reaches the platform time limit

Datasets, reference alignments, and task results are available on the HOBBIT website: `https://project-hobbit.eu/challenges/om2017/`.

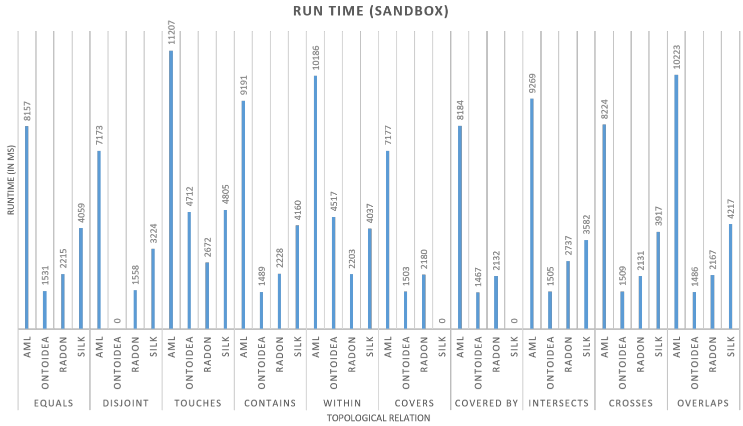


Fig. 8. HOBBIT Link Discovery Spatial Task (Sandbox)

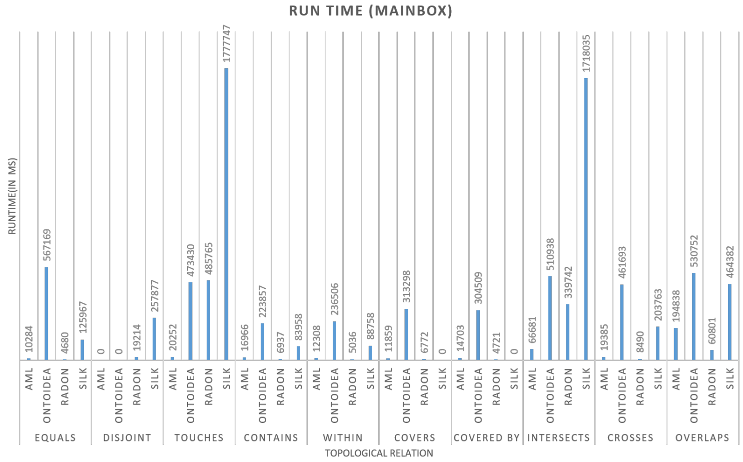


Fig. 9. HOBBIT Link Discovery Spatial Task (Mainbox)

11 Process Model Matching

In 2013 and in 2015 the community, interested in business process modeling conducted an evaluation campaign similar to the OAEI [4]. Instead of matching ontologies, the task was to match process models described in different formalisms like BPMN and Petri Nets. Within this track we offer a subset of the tasks from the Process Model Matching Contest as OAEI track by converting the process models to an ontological representation. By offering this track, we hope to gain insights in how far ontology matching systems are capable of solving the more specific problem of matching process models. This track is also motivated by the discussions at the end of the 2015 Ontology Matching workshop, where many participants showed their interest in such a track.

Table 22. Spatial Benchmark results

Relation	System	Sandbox Run Time	Mainbox Run Time
EQUALS	AML	8157	10284
	OntoIdea	1531	567169
	RADON	2215	4680
	Silk	4059	125967
DISJOINT	AML	7173	Time-out (75 min)
	OntoIdea	Not participating	
	RADON	1558	19214
	Silk	3224	257877
TOUCHES	AML	11207	20252
	OntoIdea	4712	473430
	RADON	2672	485765
	Silk	4805	1777747
CONTAINS	AML	9191	16966
	OntoIdea	1489	223857
	RADON	2228	6937
	Silk	4160	83958
WITHIN	AML	10186	12308
	OntoIdea	4517	236506
	RADON	2203	5036
	Silk	4037	88758
COVERS	AML	7177	11859
	OntoIdea	1503	313298
	RADON	2180	6772
	Silk	Not participating	
COVERED BY	AML	8184	14703
	OntoIdea	1467	304509
	RADON	2132	4721
	Silk	Not participating	
INTERSECTS	AML	9269	66681
	OntoIdea	1505	510938
	RADON	2737	339742
	Silk	3582	1718035
CROSSES	AML	8224	19385
	OntoIdea	1509	461693
	RADON	2131	8490
	Silk	3917	203763
OVERLAPS	AML	10223	194838
	OntoIdea	1486	530752
	RADON	2167	60801
	Silk	4217	464382

11.1 Experimental Settings

We used two datasets from the 2015 Process Matching Contest. The first dataset (University Admission dataset) deals with processing applications of Master students to a university. It consists of nine different process models where each describes the concrete process of a specific German university. We already used that dataset in the 2016 edition of the OAEI. The models are encoded as BPMN process models. We converted the BPMN representation of the process models to a set of assertions (ABox) using the vocabulary defined in the BPMN 2.0 ontology (TBox). The second dataset, known as the Birth Registration dataset, describes the process of registering a new born child in different countries. The process models were originally available as Petri Nets. We converted them also to an ABox in an ontological representation. For that reason the resulting matching tasks are instance matching tasks where each ABox is described by the same TBox.

For each pair of processes manually generated reference alignments are available. Typical activities within that domain are *Sending acceptance*, *Invite student for interview*, or *Wait for response*. These examples illustrate one of the main differences to the ontology matching task. The labels are usually verb-object phrases that are sometimes extended with more words. Another important difference is related to the existence of an execution order (i.e., the model is a complex sequence of activities) which can be understood as the counterpart to a type hierarchy.

Only three systems generated non-empty results when running them against our datasets. These systems are AML, LogMap, and I-Match. Note that we tried to execute all systems marked as instance matching systems. However, the other systems threw exceptions or produced empty alignments. We have collected all generated non-empty alignments. These alignments are the raw results that the following report is based on.

In our evaluation, we computed standard precision and recall, as well as the harmonic mean known as f-measure. The dataset we used consists of several test cases. We aggregated the results and present the micro average results. The gold standard we used for our first set of evaluation experiments is based on the gold standard that has also been used at the Process Model Matching Contest in 2015 [4]. We modified only some minor mistakes (resulting in changes less than 0.5 percentage points). In order to compare the results to the results obtained by the process model matching community, we present also the recomputed values of the submissions to the 2015 contest.

We extent our evaluation ("Standard" in Tables 23 and 24) by an evaluation measure that makes use of a non-binary reference alignment ("Probabilistic" in Tables 23 and 24). This probabilistic measure is based on a gold standard which is manually and independently generated by several domain experts. The number of votes of these annotators are applied as support values in the probabilistic evaluation. For a detailed discussion, please refer to [29].

Furthermore, we evaluate the matching systems via matching patterns. Therefore the matching task as well as the matcher output is automatically categorized into categories with different complexity level. We classified each alignment in one out of five categories exclusively. In this way, strength and weaknesses of the matching systems can be analysed. For more details we refer to [30].

11.2 Results

The following tables show the results of our evaluation. Participants of the Process Model Matching Contest and the OAEI 2016 edition are depicted in gray font, while this years OAEI participants are shown in black font. Note that some systems participated with a version that has not been modified with respect to its results comparing the OAEI 2016 and 2017 submission. We added only one entry for them with the label OAEI-16/17. This is only the case for the first dataset, which we have used already in 2016.

Tables 23 and 24 summarize the results of our evaluation. "P" abbreviates precision, "R" is recall, "FM" stands for f-measure and "Rk" means rank. The prefix "Pro" indicates the probabilistic versions of the precision, recall, f-measure and the associated rank. The OAEI participants are ranked on position 1, 11, 12 with an overall number of 17 systems listed in the table (when using the standard metrics). Note that AML-PM at the PMMC 2015 was a matching system that was based on a predecessor of AML participating at the OAEI 2016. The good results of AML are surprising, since we expected that matching systems specifically developed for the purpose of process model matching would outperform ontology matching systems applied to the special case of process model matching. While AML contains also components that are specifically designed for the process matching task (a flooding-like structural matching algorithm), its relevant main components are developed for ontology matching and the sub-problem of instance matching. AML and LogMap achieve the same results as in 2016. I-Match participates in 2017 for the first time. Compared to the results of the tools specialized for the problem of process model matching, the results of I-Match are still very good. There are still five systems that have in particular been designed for matching process models, which achieve worse results.

Table 23. Results of the Process Model Matching track for the University Admission dataset

Participant			**Standard**				**Probabilistic**			
System	**Contest**	**Size**	**P**	**R**	**FM**	**Rk**	**ProP**	**ProR**	**ProFM**	**Rk**
AML	OAEI-16/17	221	0.719	0.685	0.702	1	0.742	0.283	0.410	2
AML-PM	PMMC-15	579	0.269	0.672	0.385	15	0.377	0.398	0.387	4
BPLangMatch	PMMC-15	277	0.368	0.440	0.401	13	0.532	0.272	0.360	8
DKP	OAEI-16	177	0.621	0.474	0.538	8	0.686	0.219	0.333	9
DKP*	OAEI-16	150	0.680	0.440	0.534	9	0.772	0.211	0.331	10
KnoMa-Proc	PMMC-15	326	0.337	0.474	0.394	14	0.506	0.302	0.378	5
KMatch-SSS	PMMC-15	261	0.513	0.578	0.544	6	0.563	0.274	0.368	7
LogMap	OAEI-16/17	267	0.449	0.517	0.481	11	0.594	0.291	0.390	3
I-Match	OAEI-17	192	0.521	0.431	0.472	12	0.523	0.183	0.271	16
Match-SSS	PMMC-15	140	0.807	0.487	0.608	4	0.761	0.192	0.307	12
OPBOT	PMMC-15	234	0.603	0.608	0.605	5	0.648	0.258	0.369	6
pPalm-DS	PMMC-15	828	0.162	0.578	0.253	17	0.210	0.335	0.258	17
RMM-NHCM	PMMC-15	220	0.691	0.655	0.673	2	0.783	0.297	0.431	1
RMM-NLM	PMMC-15	164	0.768	0.543	0.636	3	0.681	0.197	0.306	13
RMM-SMSL	PMMC-15	262	0.511	0.578	0.543	7	0.516	0.242	0.329	11
RMM-VM2	PMMC-15	505	0.216	0.470	0.296	16	0.309	0.294	0.301	14
TripleS	PMMC-15	230	0.487	0.483	0.485	10	0.486	0.210	0.293	15

The results for the Birth Registration dataset are more interesting, because we are using this dataset in 2017 for the first time. Moreover, the dataset contains a higher amount of correspondences that are hard to find by comparing the labels on a lexical level. This results usually in a significantly lower F-measure compared to the University Admission dataset.

The results show that AML is no longer the best of all matching systems. Four systems from the process matching community achieve better results in terms of f-measure. This dataset is dominated by the OPBOT system, while AML is among a group of follow-up systems that perform still significantly better than the rest of the field. The other two systems, LogMap and I-Match, achieve close results which are slightly worse than the average results. It is interesting to see that the ranking among the three systems is the same across the two datasets.

In the probabilistic evaluation, in the University Admission dataset however, the OAEI participants gain position 2, 3, 16 respectively. LogMap rises from position 11 to 3. The (probabilistic) precision improves over-proportionally for this matcher, because LogMap generates many corre-

Table 24. Results of the Process Model Matching track for the Birth Registration dataset

Participant			Standard				Probabilistic			
System	**Contest**	**Size**	**P**	**R**	**FM**	**Rk**	**ProP**	**ProR**	**ProFM**	**Rk**
AML	OAEI-17	502	0.454	0.391	0.420	5	0.467	0.515	0.490	10
AML-PM	PMMC-15	503	0.423	0.365	0.392	7	0.513	0.505	0.509	7
BPLangMatch	PMMC-15	279	0.645	0.309	0.418	6	0.661	0.417	0.511	5
KnoMa-Proc	PMMC-15	740	0.234	0.297	0.262	15	0.224	0.437	0.296	15
KMatch-SSS	PMMC-15	185	0.800	0.254	0.385	8	0.865	0.379	0.527	4
LogMap	OAEI-17	239	0.615	0.252	0.358	11	0.834	0.411	0.551	3
I-Match	OAEI-17	188	0.734	0.237	0.358	12	0.812	0.366	0.504	8
Match-SSS	PMMC-15	128	0.922	0.202	0.332	13	0.974	0.315	0.476	11
OPBOT	PMMC-15	383	0.713	0.468	0.565	1	0.650	0.517	0.576	1
pPalm-DS	PMMC-15	490	0.502	0.422	0.459	2	0.469	0.521	0.493	9
RMM-NHCM	PMMC-15	267	0.727	0.333	0.456	3	0.781	0.443	0.565	2
RMM-NLM	PMMC-15	128	0.859	0.189	0.309	14	0.912	0.293	0.443	14
RMM-SMSL	PMMC-15	354	0.508	0.309	0.384	9	0.518	0.42	0.464	13
RMM-VM2	PMMC-15	492	0.474	0.400	0.433	4	0.454	0.48	0.466	12
TripleS	PMMC-15	266	0.613	0.280	0.384	10	0.651	0.426	0.515	6

spondences which are not included in the binary gold standard but are included in the probabilistic one. The ranking of LogMap demonstrates that a strength of the probabilistic metric lies in the broadened definition of the gold standard where weak mappings are included but softened (via the support values). In the probabilistic evaluation for the Birth Registration dataset, the three participating matchers gain ranking 3, 8 and 10. LogMap rises from rank 11 to 3 in the probabilistic evaluation. The matcher LogMap mainly identifies correspondences with high support (of which many are not included in the binary gold standard). For the matcher AML, the opposite effect can be observed. The matcher AML does not profit as much from the broadened gold standard in the probabilistic evaluation in the Birth Registration dataset compared to the other matching systems. The matchers improve their performance compared to the binary evaluation. This indicates that in the binary gold standard many reasonable alignments are missing. Thus the matchers improve their performance with the probabilistic evaluation. For details about the probabilistic metric, please refer to [29].

The results indicate that the progress made in ontology matching has also a positive impact on other related matching problems, like it is the case for process model matching. While it might require to reconfigure, adapt, and extend some parts of the ontology matching systems, such a system seems to offer a good starting point which can be turned with a reasonable amount of work into a good process matching tool. We have to emphasize that only three participants decided to apply their systems to the new track of process model matching. Thus, we have to be cautious to generalize the results we observed so far.

To allow for an in-depth analysis of the performance of the matching systems, we make use of a new evaluation method which automatically classifies the matching task into matching patterns with different attributes. The matching patterns are assigned automatically to the reference alignment, as well as to the matcher output of the three participating matchers. Then category-dependent precision, recall and f-measure are computed for each category separately. For more details please refer to [30].

Tables 25 and 26 show the results of the matching systems for each of the categories. The second column, the f-measure (FM) over all matching patterns, is given as the micro value, i.e. it

Approach	FM	Cat.			Cat. I			Cat. II			Cat. III			Cat.		
		trivial			no word iden.			one verb iden.			one word iden.			misc		
		[44.3%][103]			[29.3%][68]			[11.6%][27]			[7.3%][17]			[7.3%][17]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML	.702	.890	.942	.915	.953	.603	.739	.833	.185	.303	.667	.353	.462	.167	.529	.254
I-Match	.472	.907	.942	.924	–	–	–	.400	.074	.125	–	–	–	.500	.059	.105
LogMap	.481	.894	.981	.935	–	–	–	.500	.148	.229	.133	.353	.194	.089	.529	.153

Table 25. Results assigned to matching patterns of University Admission dataset

Approach	FM	Cat.			Cat. I			Cat. II			Cat. III			Cat.		
		trivial			no word iden.			one verb iden.			one word iden.			misc		
		[4.5%][26]			[74.9%][437]			[1.5%][9]			[9.9%][58]			[9.1%][53]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML	.420	.759	.846	.800	.427	.364	.393	.133	.222	.167	.438	.362	.396	.632	.453	.527
I-Match	.358	.950	.731	.826	.746	.236	.358	.667	.222	.333	.400	.103	.164	.667	.151	.246
LogMap	.358	.339	.731	.463	.726	.261	.384	–	–	–	.357	.086	.139	.818	.170	.281

Table 26. Results assigned to matching patterns of Birth Registration dataset

is computed over all test cases. The remaining columns provide the category-dependent precision (cP), recall (cR) and f-measure (cFM) for each matcher in each category. cP, cR and cFM are macro values, independently computed for each category. Moreover, for each category, the tables contain in the heading the fraction of correspondences from the whole data set as well as the total number of correspondences of a category in the reference alignment. Cat. I contains alignments which have no word in common (syntactically). It can be observed that for the University Admission dataset it is sufficient to identify mainly trivial correspondences. I-Match and LogMap do not compute any alignments of the most complex category ("Cat. I"). However, AML has a very high performance for "Cat. I". In the Birth Registration dataset the fraction of trivial alignments is very low. The most dominant category is "Cat. I". Therefore, it is not sufficient to focus on the identification of trivial alignments. In contrast to the University Admission datatset, the matchers compute reasonable alignments from "Cat. I" in the Birth Registration dataset. The low performance of the three matchers for "Cat. trivial" in the Birth Registration dataset indicates mistakes in the binary gold standard.

11.3 Conclusions

In 2016 we organized the Process Model Matching track for the first time. Our evaluation effort was motivated by the idea that Ontology Matching methods and techniques can also be used in the related field of Process Model Matching. For that reason we converted one (and in 2017 two) of the most prominent Process Model matching test datasets into an ontological representation. The resulting matching problems are instance matching tasks.

While we were aware that an instance matching system will not be able to exploit the sequential aspects of the given process models out of the box, we expected lexical components to generate results that are already on an acceptable level. Even though some of the systems generated very good results, overall only a few of the systems participating at the OAEI were capable of generating any results for our test cases. We still do not fully understand the reasons for this outcome.

In order to facilitate the evaluation process for participants which cannot evaluate their matchers with SEALS, we developed a web-based evaluation platform[21] to potentially increase the number of participants. This platform was intended to be used by potential participants from the process matching community that are not interested in an OAEI participation, which is tailored for ontology matching systems. Within this platform, participants are able to select one or multiple gold standards for one of the datasets and subsequently upload their corresponding matcher results. Afterwards, the participants are able to select from a variety of different metrics including not only different types of precision, recall and f-measure but also general statistics for the generated output. Unfortunately, no further matching systems participated via the platform.

The participation rate indicates that only a limited number of participants is interested in process model matching. For that reason we will not offer a third edition of this track in 2018.

12 Statistical analysis

The traditional evaluation carried out in the OAEI tracks consists simply of comparing and ranking systems based on performance scores such as F-measure. In the case of tracks with multiple datasets, performance scores are averaged for all datasets, and the systems are compared accordingly. While performance scores enable us to gage the performance of matching systems individually, they are insufficient for drawing statistically meaningful comparisons between systems.

In the interest of providing a more in-depth comparison of the matching systems that participated in this year's competition, this section presents an analysis based on statistical inference.

12.1 Methods

For one-dataset comparisons, we use McNemar's test. This test takes as input the alignments produced by two matching systems plus the reference alignment, and produces as output an indicator which shows if either system is better than the other or whether they are approximately the same. This method of comparison does not need a particular performance score to be determined beforehand. Further, the comparison is not solely based on the juxtaposition of two scalars, but rather, it is substantiated by the statistical evidence (null hypothesis testing). Two variants of McNemar's test were considered: one where false correspondences were ignored so that the comparison was predicated only on the correct correspondences found by matching systems; and another where both correct and false correspondences were considered, meaning that systems were compared based on the full alignment they generated. A directed graph can be used to visualize the outcome of the test. Interested readers are referred to [34] for more details about the utilization of this methodology.

For comparisons over multiple datasets, we used the Friedman test with the corresponding post-hoc procedure for comparison. This test requires the specification of one performance score. The outcome of the test can be visualized by critical difference (CD) diagrams.

Since the comparisons between matching systems are done pairwise, it is necessary to correct the statistics for multiple testing. We used the Bergmann correction method to control the family-wise error rate in all tests.

[21] http://alkmaar.informatik.uni-mannheim.de/pmmc

12.2 Results

Anatomy track In this year's competition, 11 systems participate in the anatomy track. However, the alignments of the LogMap family could not be parsed by the Alignment API, so we had to leave them out from the comparative analysis for this track.

Figure 10 shows the directed graph with the outcome of McNemar's test over participatory systems when the false correspondences are not taken into account. Figure 11 shows the corresponding result when all correspondences are considered. The nodes in these graphs are the systems and a directed edge $A \rightarrow B$ indicates the superiority of A over B. If there is no such an edge between any two systems, then they are claimed to be more or less equivalent.

According to these figures, AML is the best system and Wiki3 and ALIN are the bottom ones, from both perspectives. There are two differences between the two approaches to conducting the test. SANOM outperforms KEPLER when the false correspondences are not considered, and KEPLER is better than SANOM if wrong correspondences are taken into account. It means that SANOM discovers more correct correspondences than KEPLER, but also more false correspondences. A similar pattern holds for the comparison of POMAP and YAM-BIO. Interestingly, no systems are declared to be equivalent, so the outcome of McNemar's test is similar to a ranking scheme.

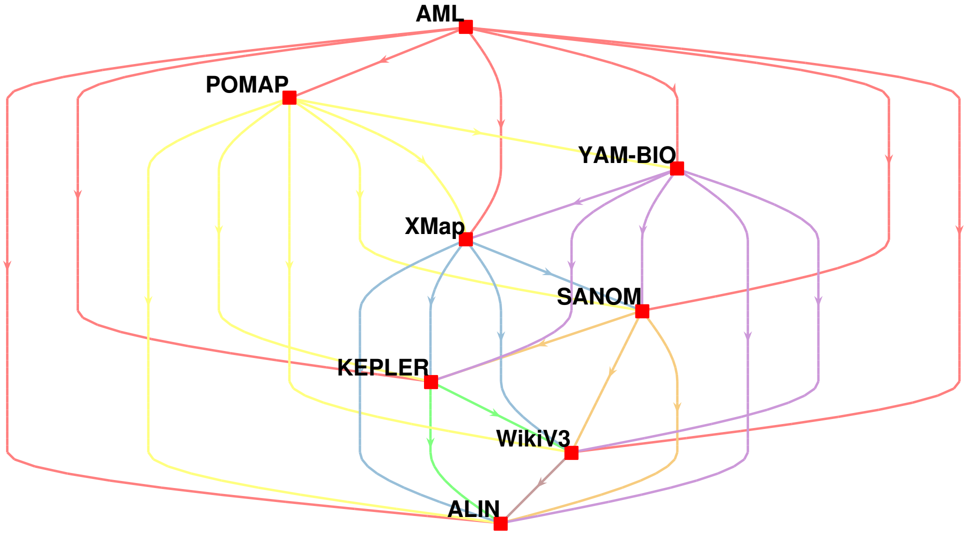


Fig. 10. Comparison of alignment systems participated in OAEI 2017 on the anatomy track while the false correspondences are not considered.

Conference track This track consists of 21 small matching tasks between 7 different ontologies. Three different types of matching are considered: (i) M1: only matching the classes; (i) M2: only matching the properties; (ii) M3: matching both classes and properties. The reference alignment has also three different variants. Hence, there are nine different modes of evaluating systems, based on the type of matching and the type of reference alignment. The Friedman test

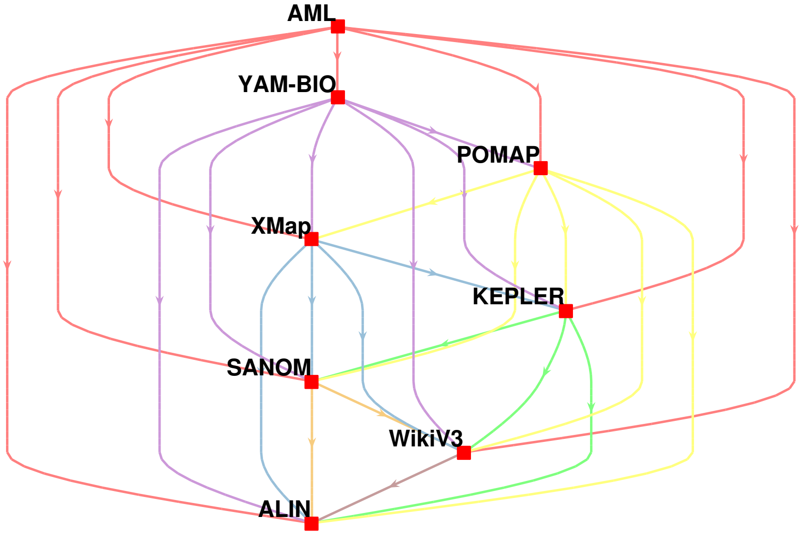


Fig. 11. Comparison of alignment systems participated in OAEI 2017 on the anatomy track while the false correspondences are taken into account.

was applied considering the F-measure of the systems on each of the 21 tasks for each of the evaluation modes.

Figure 12 shows the CD diagram of the systems that participated in this track. In this figure, the x axis is the average rank obtained by the Friedman test, and the systems with the same performance are connected to each other by the red lines. The lower the average rank in the CD plot, the better the performance of the system.

The CD diagram for this track provides little information and insight about the difference between systems, likely due to the small sample size for the comparison (systems produce only between 90 and 240 correspondences in total in this track). What is readily seen from this plot is the superiority of AML, LogMap, and XMap and the poor performance of ALIN, SANOM, and POMap.

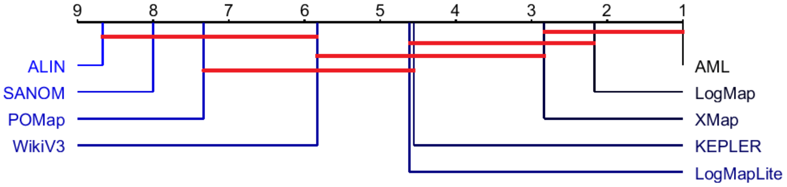


Fig. 12. Comparison of alignment systems participated in OAEI 2017 on the Conference track. The x-axis is the average rank of each system obtained by the Friedman test. Systems which are not significantly different from each other are connected by the red lines.

LargeBio track This track consists six matching tasks of large size. The Friedman test was applied to the F-measure obtained by each system over each alignment task. Figure 13 shows the corresponding CD diagram for this track. According to this plot, the group containing AML, XMap, YAM-BIO, LogMap, and LogMapBio are the best systems, and POMAP, SANOM, and KEPLER are the systems with lackluster performance in this track.

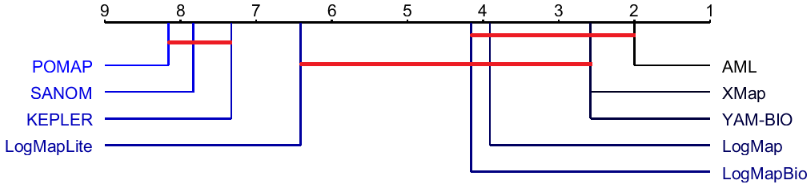


Fig. 13. Comparison of alignment systems participated in OAEI 2017 on the LargeBio track. The x-axis is the average rank of each system obtained by the Friedman test. Systems which are not significantly different from each other are connected by the red lines.

Multifarm track This track involves 55 matching tasks with ontologies from different languages. The Friedman test was applied to the F-measure obtained by each system over each task. The CD diagram depicting the outcome of the test is shown in Figure 14.

According to this graph, AML is exclusively the best alignment system in this track. LogMap, CroLOM, and KEPLER perform equally better than the remaining systems. At the other extreme, LogMapLite, XMap, and SANOM show a poor performance in this track, while WikiV3 ranks in between the two trios.

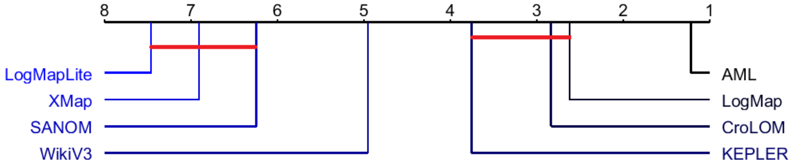


Fig. 14. Comparison of alignment systems participated in OAEI 2017 on the Conference track. The x-axis is the average rank of each system obtained by the Friedman test. Systems which are not significantly different from each other are connected by the red lines.

13 Lesson learned and suggestions

The lessons learned from running OAEI 2017 were the following:

A) Like last year, this year we requested tool registration in June and preliminary submission of wrapped systems by the end of July, but were more strict in its enforcement. As a result, we recorded the smallest number of errors and incompatibilities with the SEALS client during the evaluation phase in recent OAEI editions.

B) As has been the trend, some system developers struggled to get their systems working with the SEALS client, mostly due to incompatible versions of libraries. While participation on the new HOBBIT track was relatively low due to the novelty of the HOBBIT platform and the short deadline for systems to adapt to it, the solution of using Docker containers to wrap systems seems promising, and we are already looking into phasing out the SEALS client in favour of the HOBBIT platform.

C) While the number of participants this year was similar to that of recent years, their distribution through the tracks was uneven. The expressive ontologies tracks had no shortage of participants, and still a fair number participated in the more specialized multifarm track. However, participation in the interactive matching track and in the three instance matching tracks (process model, instance, and hobbit) was underwhelming. The latter is puzzling considering the prize sponsored by IBM Research for the system with the best performance across the instance matching tracks. Granted, the division of instance matching tracks between the SEALS client and the HOBBIT platform did not help their cause, as of the 7 total systems that participated in instance matching tasks, only 2 made both a SEALS and a HOBBIT submission. Nevertheless, the division between "traditional" ontology matching and instance matching is readily apparent, as only 2 systems have participated in both track families.

D) In previous years we identified the need for considering non-binary forms of evaluation, namely in cases where there is uncertainty about some of the reference mappings. A first non-binary evaluation type was implemented in the Conference track in 2015, followed by Disease and Phenotype, and Process Model in 2016. This year, we have introduced statistical tests to compare matching systems, an analysis that was carried out on the results of 4 tracks. This approach provides more insights into the comparative performance of systems as well as more statistical rigour, and thus we hope that it can be expanded and fully integrated into the OAEI tracks in future editions.

The lessons learned in the various OAEI 2017 track were the following:

conference: Since there have been no improvement in matchers performance this year from the perspective of performed evaluation modalities we will consider to add or replace existing evaluation modalities for future editions of OAEI to help disclose further matchers characteristics.

largebio: While the current reference alignments, with incoherence-causing mappings flagged as uncertain, make the evaluation fair to all systems, they are only a compromise solution, not an ideal one. Thus, we should aim for manually repairing and validating the reference alignments for future editions.

phenotype: This track attracted a similar level of participation this year compared to last, despite no cash prize, which demonstrates its intrinsic value and interest among the community of ontology matching algorithm developers.

interactive: This track's participation has remained low, as most systems participating in OAEI opt to focus exclusively on fully automatic matching. We hope to draw more participants to this track in the future and will continue to expand it so as to better approximate real user interactions.

process model: The results of the Process Model track have shown that the participating ontology matching systems are capable of generating good results for the specific problem of process model matching, even though few were able to exploit the sequential aspects of the process models. Even though we offered an alternative evaluation process for participants which cannot evaluate their matchers with SEALS, this alternative failed to attract further participants. The low participation rate in this track indicates that only a limited number of participants is interested in process model matching. For that reason we will not offer a third edition of this track in 2018.

instance: In order to attract more instance matching systems to participate in value semantics (val-sem), value structure (val-struct), and value structure semantics (val-struct-sem) tasks, we need to produce benchmarks that have fewer instances (in the order of 10000), of the same type (in our benchmark we asked systems to compare instances of different types). To balance those aspects, we must then produce benchmarks that contain more complex transformations.

14 Conclusions

The OAEI 2017 saw the same number of participants as in recent years, with a healthy mix of new and returning systems. While last year we posited that new participants were drawn by the allure of prize money in the new Disease and Phenotype track, the evidence this year seems to contradict it. On the one hand, participation in Disease and Phenotype remain high this year despite no prize money. On the other hand, the prize money on offer for performance in instance matching did not attract many participants to those tracks. Nevertheless, the fact that there continues to be corporate interest in ontology matching to the point of offering prize money bodes well for the future of the OAEI.

Like last year, judging from the repeated tracks, there has been no substantial progress to the state of the art in ontology matching overall this year:

- There was no noticeable improvement with regard to system run times.
- There were few improvements with regard to F-measure, with the top results in most tracks remaining the same.
- There was no significant progress with regard to the ability of matching systems to handle large ontologies and datasets, either in traditional ontology matching or in instance matching.
- There was no progress with regard to alignment repair systems, with only a few returning systems employing them.

This conclusion may be due to a plateau being reached by matching systems in some tracks, and investing in improving results further would bring diminishing returns. However, it is also the case that long-term participants tend to focus more on the new datasets and tracks on offer than on improving in repeated tracks. Given the variety of tracks on offer, it is difficult for system developers to aim at improving across all tracks each year.

Most of the participants have provided a description of their systems and their experience in the evaluation. These OAEI papers, like the present one, have not been peer reviewed. However, they are full contributions to this evaluation exercise and reflect the hard work and clever insight people put into the development of participating systems. Reading the papers of the participants should help people involved in ontology matching find out what makes these algorithms work and what could be improved.

The Ontology Alignment Evaluation Initiative will strive to remain a reference to the ontology matching community by improving both the test cases and the testing methodology to better reflect actual needs, as well as to promote progress in this field [43]. More information can be found at:

`http://oaei.ontologymatching.org`.

Acknowledgements

We warmly thank the participants of this campaign. We know that they have worked hard to have their matching tools executable in time and they provided useful reports on their experience. The best way to learn about the results remains to read the papers that follow.

We would also like to thank IBM Research for sponsoring the instance matching tracks by offering prize money for the best performing systems.

We are grateful to the Universidad Politécnica de Madrid (UPM), especially to Nandana Mihindukulasooriya and Asunción Gómez Pérez, for moving, setting up and providing the necessary infrastructure to run the SEALS repositories.

We are also grateful to Martin Ringwald and Terry Hayamizu for providing the reference alignment for the anatomy ontologies and thank Elena Beisswanger for her thorough support on improving the quality of the data set.

We thank Khiat Abderrahmane for his support in the Arabic data set and Catherine Comparot for her feedback and support in the MultiFarm test case.

We also thank for their support the other members of the Ontology Alignment Evaluation Initiative steering committee: Yannis Kalfoglou (Ricoh laboratories, UK), Miklos Nagy (The Open University (UK), Natasha Noy (Google Inc., USA), Yuzhong Qu (Southeast University, CN), York Sure (Leibniz Gemeinschaft, DE), Jie Tang (Tsinghua University, CN), Heiner Stuckenschmidt (Mannheim Universität, DE), George Vouros (University of the Aegean, GR).

Michelle Cheatham has been supported by the National Science Foundation award ICER-1440202 "EarthCube Building Blocks: Collaborative Proposal: GeoLink".

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

Daniel Faria was supported by the ELIXIR-EXCELERATE project (INFRADEV-3-2015).

Ernesto Jimenez-Ruiz has also been partially supported by the BIGMED project (IKT 259055), the HealthInsight project (IKT 247784), the SIRIUS Centre for Scalable Data Access (Research Council of Norway, project no.: 237889).

Catia Pesquita was supported by the FCT through the LASIGE Strategic Project (UID/CEC/00408/2013) and the research grant PTDC/EEI-ESS/4633/2014.

References

1. Manel Achichi, Rodolphe Bailly, Cécile Cecconi, Marie Destandau, Konstantin Todorov, and Raphaël Troncy. Doremus: Doing reusable musical data. In *ISWC PD: International Semantic Web Conference Posters and Demos*, 2015.
2. Manel Achichi, Michelle Cheatham, Zlatan Dragisic, Jerome Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Ian Harrow, Valentina Ivanova, Ernesto Jiménez-Ruiz, Elena Kuss, Patrick Lambrix, Henrik Leopold, Huanyu Li, Christian Meilicke, Stefano Montanelli, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Heiner Stuckenschmidt, Konstantin Todorov, Cássia Trojahn, and Ondrej Zamazal. Results of the ontology alignment evaluation initiative 2016. In *Proc. 11th ISWC ontology matching workshop (OM), Kobe (JP)*, pages 73–129, 2016.
3. José Luis Aguirre, Bernardo Cuenca Grau, Kai Eckert, Jérôme Euzenat, Alfio Ferrara, Robert Willem van Hague, Laura Hollink, Ernesto Jiménez-Ruiz, Christian Meilicke, Andriy Nikolov, Dominique Ritze, François Scharffe, Pavel Shvaiko, Ondrej Sváb-Zamazal, Cássia Trojahn, and Benjamin Zapilko. Results of the ontology alignment evaluation initiative 2012. In *Proc. 7th ISWC ontology matching workshop (OM), Boston (MA, US)*, pages 73–115, 2012.

4. Gonçalo Antunes, Marzieh Bakhshandeh, José Borbinha, João Cardoso, Sharam Dadashnia, Chiara Di Francescomarino, Mauro Dragoni, Peter Fettke, Avigdor Gal, Chiara Ghidini, Philip Hake, Abderrahmane Khiat, Christopher Klinkmüller, Elena Kuss, Henrik Leopold, Peter Loos, Christian Meilicke, Tim Niesen, Catia Pesquita, Timo Péus, Andreas Schoknecht, Eitam Sheetrit, Andreas Sonntag, Heiner Stuckenschmidt, Tom Thaler, Ingo Weber, and Matthias Weidlich. The process model matching contest 2015. In *6th EMISA Workshop*, pages 127–155, 2015.
5. Benhamin Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Proc. K-Cap Workshop on Integrating Ontologies*, Banff (Canada), 2005.
6. Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.
7. Caterina Caracciolo, Jérôme Euzenat, Laura Hollink, Ryutaro Ichise, Antoine Isaac, Véronique Malaisé, Christian Meilicke, Juan Pane, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, and Vojtech Svátek. Results of the ontology alignment evaluation initiative 2008. In *Proc. 3rd ISWC ontology matching workshop (OM), Karlsruhe (DE)*, pages 73–120, 2008.
8. Michelle Cheatham, Zlatan Dragisic, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, et al. Results of the ontology alignment evaluation initiative 2015. In *Proc. 10th ISWC ontology matching workshop (OM), Bethlehem (PA, US)*, pages 60–115, 2015.
9. Bernardo Cuenca Grau, Zlatan Dragisic, Kai Eckert, Jérôme Euzenat, Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, Andreas Oskar Kempf, Patrick Lambrix, Andriy Nikolov, Heiko Paulheim, Dominique Ritze, François Scharffe, Pavel Shvaiko, Cássia Trojahn dos Santos, and Ondrej Zamazal. Results of the ontology alignment evaluation initiative 2013. In Pavel Shvaiko, Jérôme Euzenat, Kavitha Srinivas, Ming Mao, and Ernesto Jiménez-Ruiz, editors, *Proc. 8th ISWC ontology matching workshop (OM), Sydney (NSW, AU)*, pages 61–100, 2013.
10. Jim Dabrowski and Ethan V. Munson. 40 years of searching for the best computer system response time. *Interacting with Computers*, 23(5):555–564, 2011.
11. Jérôme David, Jérôme Euzenat, François Scharffe, and Cássia Trojahn dos Santos. The alignment API 4.0. *Semantic web journal*, 2(1):3–10, 2011.
12. Cássia Trojahn dos Santos, Bo Fu, Ondrej Zamazal, and Dominique Ritze. State-of-the-art in multilingual and cross-lingual ontology matching. In *Towards the Multilingual Semantic Web, Principles, Methods and Applications*, pages 119–135. 2014.
13. Zlatan Dragisic, Kai Eckert, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, Andreas Oskar Kempf, Patrick Lambrix, Stefano Montanelli, Heiko Paulheim, Dominique Ritze, Pavel Shvaiko, Alessandro Solimando, Cássia Trojahn dos Santos, Ondrej Zamazal, and Bernardo Cuenca Grau. Results of the ontology alignment evaluation initiative 2014. In *Proc. 9th ISWC ontology matching workshop (OM), Riva del Garda (IT)*, pages 61–104, 2014.
14. Zlatan Dragisic, Valentina Ivanova, Patrick Lambrix, Daniel Faria, Ernesto Jiménez-Ruiz, and Catia Pesquita. User validation in ontology alignment. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I*, pages 200–217, 2016.
15. Zlatan Dragisic, Valentina Ivanova, Huanyu Li, and Patrick Lambrix. Experiences from the anatomy track in the ontology alignment evaluation initiative. *Journal of Biomedical Semantics*, 2017.
16. Jérôme Euzenat, Alfio Ferrara, Laura Hollink, Antoine Isaac, Cliff Joslyn, Véronique Malaisé, Christian Meilicke, Andriy Nikolov, Juan Pane, Marta Sabou, François Scharffe, Pavel Shvaiko, Vassilis Spiliopoulos, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, Cássia Trojahn dos Santos, George Vouros, and Shenghui Wang. Results of

the ontology alignment evaluation initiative 2009. In *Proc. 4th ISWC ontology matching workshop (OM), Chantilly (VA, US)*, pages 73–126, 2009.

17. Jérôme Euzenat, Alfio Ferrara, Christian Meilicke, Andriy Nikolov, Juan Pane, François Scharffe, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, and Cássia Trojahn dos Santos. Results of the ontology alignment evaluation initiative 2010. In *Proc. 5th ISWC ontology matching workshop (OM), Shanghai (CN)*, pages 85–117, 2010.
18. Jérôme Euzenat, Alfio Ferrara, Robert Willem van Hague, Laura Hollink, Christian Meilicke, Andriy Nikolov, François Scharffe, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, and Cássia Trojahn dos Santos. Results of the ontology alignment evaluation initiative 2011. In *Proc. 6th ISWC ontology matching workshop (OM), Bonn (DE)*, pages 85–110, 2011.
19. Jérôme Euzenat, Antoine Isaac, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Svab, Vojtech Svatek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative 2007. In *Proc. 2nd ISWC ontology matching workshop (OM), Busan (KR)*, pages 96–132, 2007.
20. Jérôme Euzenat, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, and Cássia Trojahn dos Santos. Ontology alignment evaluation initiative: six years of experience. *Journal on Data Semantics*, XV:158–192, 2011.
21. Jérôme Euzenat, Malgorzata Mochol, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Svab, Vojtech Svatek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative 2006. In *Proc. 1st ISWC ontology matching workshop (OM), Athens (GA, US)*, pages 73–95, 2006.
22. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition, 2013.
23. Daniel Faria, Ernesto Jiménez-Ruiz, Catia Pesquita, Emanuel Santos, and Francisco M. Couto. Towards Annotating Potential Incoherences in BioPortal Mappings. In *13th International Semantic Web Conference*, volume 8797 of *Lecture Notes in Computer Science*, pages 17–32. Springer, 2014.
24. Ian Harrow, Ernesto Jiménez-Ruiz, Andrea Splendiani, Martin Romacker, Peter Woollard, Scott Markel, Yasmin Alam-Faruque, Martin Koch, James Malone, and Arild Waaler. Matching Disease and Phenotype Ontologies in the Ontology Alignment Evaluation Initiative. *Journal of Biomedical Semantics*, 2018.
25. Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. LogMap: Logic-based and scalable ontology matching. In *Proc. 10th International Semantic Web Conference (ISWC), Bonn (DE)*, pages 273–288, 2011.
26. Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, Ian Horrocks, and Rafael Berlanga. Logic-based assessment of the compatibility of UMLS ontology sources. *J. Biomed. Sem.*, 2, 2011.
27. Ernesto Jiménez-Ruiz, Christian Meilicke, Bernardo Cuenca Grau, and Ian Horrocks. Evaluating mapping repair systems with large biomedical ontologies. In *Proc. 26th Description Logics Workshop*, 2013.
28. Yevgeny Kazakov, Markus Krötzsch, and Frantisek Simancik. Concurrent classification of EL ontologies. In *Proc. 10th International Semantic Web Conference (ISWC), Bonn (DE)*, pages 305–320, 2011.
29. Elena Kuss, Henrik Leopold, Han Van der Aa, Heiner Stuckenschmidt, and Hajo A Reijers. Probabilistic evaluation of process model matching techniques. In *Conceptual Modeling: 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016, Proceedings 35*, pages 279–292. Springer, 2016.
30. Elena Kuss and Heiner Stuckenschmidt. Automatic classification to matching patterns for process model matching evaluation. In *Proceedings of the ER Forum 2017 and the ER 2017 Demo Track co-located with the 36th International Conference on Conceptual Modelling (ER 2017), Valencia, Spain, - November 6-9, 2017.*, pages 292–305, 2017.

31. Pasquale Lisena, Manel Achichi, Eva Fernández, Konstantin Todorov, and Raphaël Troncy. Exploring linked classical music catalogs with overture. In *ISWC PD: International Semantic Web Conference Posters and Demos*, 2016.
32. Christian Meilicke. *Alignment Incoherence in Ontology Matching*. PhD thesis, University Mannheim, 2011.
33. Christian Meilicke, Raúl García Castro, Frederico Freitas, Willem Robert van Hage, Elena Montiel-Ponsoda, Ryan Ribeiro de Azevedo, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, Andrei Tamilin, Cássia Trojahn, and Shenghui Wang. MultiFarm: A benchmark for multilingual ontology matching. *Journal of web semantics*, 15(3):62–68, 2012.
34. Majid Mohammadi, Amir Ahooye Atashin, Wout Hofman, and Yaohua Tan. Comparison of ontology alignment algorithms across single matching task via the McNemar test. *arXiv*, arXiv:1704.00045.
35. Boris Motik, Rob Shearer, and Ian Horrocks. Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009.
36. Heiko Paulheim, Sven Hertling, and Dominique Ritze. Towards evaluating interactive ontology matching tools. In *Proc. 10th Extended Semantic Web Conference (ESWC), Montpellier (FR)*, pages 31–45, 2013.
37. Catia Pesquita, Daniel Faria, Emanuel Santos, and Francisco Couto. To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. In *Proc. 8th ISWC ontology matching workshop (OM), Sydney (AU)*, page this volume, 2013.
38. Manuel Salvadores, Paul R. Alexander, Mark A. Musen, and Natalya Fridman Noy. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. *Semantic Web*, 4(3):277–284, 2013.
39. Emanuel Santos, Daniel Faria, Catia Pesquita, and Francisco Couto. Ontology alignment repair through modularization and confidence-based heuristics. *CoRR*, abs/1307.5322, 2013.
40. T. Saveta, E. Daskalaki, G. Flouris, I. Fundulaki, M. Herschel, and A.-C. Ngonga Ngomo. Pushing the limits of instance matching systems: A semantics-aware benchmark for linked data. In *WWW, Companion Volume*, 2015.
41. Tzanina Saveta, Evangelia Daskalaki, Giorgos Flouris, Irini Fundulaki, Melanie Herschel, and Axel-Cyrille Ngonga Ngomo. Lance: Piercing to the heart of instance matching tools. In *International Semantic Web Conference*, pages 375–391. Springer, 2015.
42. Mohamed Ahmed Sherif, Kevin Dreßler, Panayiotis Smeros, and Axel-Cyrille Ngonga Ngomo. RADON - Rapid Discovery of Topological Relations. In *Proceedings of The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.
43. Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):158–176, 2013.
44. Alessandro Solimando, Ernesto Jiménez-Ruiz, and Giovanna Guerrini. Detecting and correcting conservativity principle violations in ontology-to-ontology mappings. In *The Semantic Web–ISWC 2014*, pages 1–16. Springer, 2014.
45. Alessandro Solimando, Ernesto Jimenez-Ruiz, and Giovanna Guerrini. Minimizing conservativity violations in ontology alignments: Algorithms and evaluation. *Knowledge and Information Systems*, 2016.
46. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proc. ISWC Workshop on Evaluation of Ontology-based Tools (EON), Hiroshima (JP)*, 2004.

Montpellier, Dayton, Linköping, Grenoble, Lisboa,
Milano, Heraklion, Kent, Oslo, Mannheim, Amsterdam,
Delft, Trento, Toulouse, Prague
December 2017