Results of the Ontology Alignment Evaluation Initiative 2018*

Alsayed Algergawy[1], Michelle Cheatham[2], Daniel Faria[3], Alfio Ferrara[4], Irini Fundulaki[5], Ian Harrow[6], Sven Hertling[7], Ernesto Jiménez-Ruiz[8,9], Naouel Karam[10], Abderrahmane Khiat[11], Patrick Lambrix[12], Huanyu Li[12], Stefano Montanelli[4], Heiko Paulheim[7], Catia Pesquita[13], Tzanina Saveta[5], Daniela Schmidt[14], Pavel Shvaiko[15], Andrea Splendiani[6], Elodie Thiéblin[16], Cássia Trojahn[16], Jana Vataščinová[17], Ondřej Zamazal[17], and Lu Zhou[2]

[1] Friedrich Schiller University Jena, Germany
alsayed.algergawy@uni-jena.de
[2] Data Semantics (DaSe) Laboratory, Wright State University, USA
{michelle.cheatham, zhou.34}@wright.edu
[3] Instituto Gulbenkian de Ciência, Lisbon, Portugal
dfaria@igc.gulbenkian.pt
[4] Università degli studi di Milano, Italy
{alfio.ferrara,stefano.montanelli}@unimi.it
[5] Institute of Computer Science-FORTH, Heraklion, Greece
{jsaveta,fundul}@ics.forth.gr
[6] Pistoia Alliance Inc., USA
{ian.harrow,andrea.splendiani}@pistoiaalliance.org
[7] University of Mannheim, Germany
{sven,heiko}@informatik.uni-mannheim.de
[8] Department of Informatics, University of Oslo, Norway
ernestoj@ifi.uio.no
[9] The Alan Turing Institute, London, UK
ejimenez-ruiz@turing.ac.uk
[10] Fraunhofer FOKUS, Berlin, Germany
naouel.karam@fokus.fraunhofer.de
[11] Freie Universität Berlin, Germany
abderrahmane.khiat@fu-berlin.de
[12] Linköping University & Swedish e-Science Research Center, Linköping, Sweden
{patrick.lambrix,huanyu.li}@liu.se
[13] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
cpesquita@di.fc.ul.pt
[14] Pontifical Catholic University of Rio Grande do Sul, Brazil
daniela.schmidt@acad.pucrs.br
[15] TasLab, Trentino Digitale SpA, Trento, Italy
pavel.shvaiko@tndigit.it
[16] IRIT & Université Toulouse II, Toulouse, France
{cassia.trojahn,elodie.thieblin}@irit.fr
[17] University of Economics, Prague, Czech Republic
ondrej.zamazal@vse.cz

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

Abstract. 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 2018 campaign offered 12 tracks with 23 test cases, and was attended by 19 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 [18, 20]. 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, 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) [45]. 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) [4]. From 2006 until the present, the OAEI campaigns were held at the Ontology Matching workshop, collocated with ISWC [1–3, 6–8, 11, 14–17, 19], which this year took place in Monterey, CA, USA[2].

Since 2011, we have been using an environment for automatically processing evaluations (§2.1) 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. Since OAEI 2017, a novel evaluation environment called HOBBIT (§2.1) was adopted for the HOBBIT Link Discovery track, and later extended to enable the evaluation of other tracks. Some tracks are run exclusively through SEALS and others through HOBBIT, but several allow participants to choose the platform they prefer.

This paper synthesizes the 2018 evaluation campaign and introduces the results provided in the papers of the participants. The remainder of the paper is organized as follows: in §2, we present the overall evaluation methodology; in §3 we present the tracks and datasets; in §4 we present and discuss the results; and finally, §5 concludes the paper.

[1] http://oaei.ontologymatching.org
[2] http://om2018.ontologymatching.org
[3] http://www.seals-project.eu

2 Methodology

2.1 Evaluation platforms

The OAEI evaluation was carried out in one of two alternative platforms: the SEALS client or the HOBBIT platform. Both have the goal of ensuring reproducibility and comparability of the results across matching systems.

The **SEALS client** was developed in 2011. It is a Java-based command line interface for ontology matching evaluation, which requires system developers 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 is provided to the participants, describing how to wrap a tool and how to run a full evaluation locally.

The **HOBBIT platform**[4] was introduced in 2017. It is a web interface for linked data and ontology matching evaluation, which requires systems to be wrapped inside docker containers and include a SystemAdapter class, then being uploaded into the HOBBIT platform [31].

Both platforms compute the standard evaluation metrics against the reference alignments: precision, recall and F-measure. In test cases where different evaluation modalities are required, evaluation was carried out *a posteriori*, using the alignments produced by the matching systems.

2.2 OAEI campaign phases

As in previous years, the OAEI 2018 campaign was divided into three phases: preparatory, execution, and evaluation.

In the **preparatory phase**, the test cases were provided to participants in an initial assessment period between June 15^{th} and July 15^{th}, 2018. The goal of this phase is to ensure that the test cases make sense to participants, and give them the opportunity to provide feedback to organizers on the test case as well as potentially report errors. At the end of this phase, the final test base was frozen and released.

During the ensuing **execution phase**, participants test and potentially develop their matching systems to automatically match the test cases. Participants can self-evaluate their results either by comparing their output with the reference alignments or by using either of the evaluation platforms. They can tune their systems with respect to the non-blind evaluation as long as they respect the rules of the OAEI. Participants were required to register their systems and make a preliminary evaluation by July 31^{st}. The execution phase was terminated on September 9^{th}, 2018, at which date participants had to submit the (near) final versions of their systems (SEALS-wrapped and/or HOBBIT-wrapped).

During the **evaluation phase**, systems were evaluated by all track organizers. In case minor problems were found during the initial stages of this phase, they were reported to developers, who were given the opportunity to fix and resubmit their systems. Initial results were provided directly to the participants, whereas final results for most tracks were published on the respective pages of the OAEI website by October 8^{th}.

[4] `https://project-hobbit.eu/outcomes/hobbit-platform/`

3 Tracks and test cases

This year's OAEI campaign consisted of 12 tracks gathering 23 test cases, all of which were based on OWL ontologies. They can be grouped into:

- Schema matching tracks, which have as objective matching ontology classes and/or properties.
- Instance Matching tracks, which have as objective matching ontology instances.
- Instance and Schema Matching tracks, which involve both of the above.
- Complex Matching tracks, which have as objective finding complex correspondences between ontology entities.
- Interactive tracks, which simulate user interaction to enable the benchmarking of interactive matching algorithms.

The tracks are summarized in Table 1.

Table 1. Characteristics of the OAEI tracks.

Track	Test Cases (Tasks)	Relations	Confidence	Evaluation	Languages	Platform
Schema Matching						
Anatomy	1	=	[0 1]	open	EN	SEALS
Biodiversity & Ecology	2	=	[0 1]	open	EN	SEALS
Conference	1 (21)	=, <=	[0 1]	open+blind	EN	SEALS
Disease & Phenotype	2	=, <=	[0 1]	open+blind	EN	SEALS
Large Biomedical ontologies	6	=	[0 1]	open	EN	both
Multifarm	2 (2695)	=	[0 1]	open+blind	AR, CZ, CN, DE, EN, ES, FR, IT, NL, RU, PT	SEALS
Instance Matching						
IIMB	1	=	[0 1]	open+blind	EN	SEALS
Link Discovery	2 (9)	=	[0 1]	open	EN	HOBBIT
SPIMBENCH	2	=	[0 1]	open+blind	EN	HOBBIT
Instance and Schema Matching						
Knowledge Graph	9	=	[0 1]	open	EN	both
Interactive Matching						
Interactive	2 (22)	=, <=	[0 1]	open	EN	SEALS
Complex Matching						
Complex	4	=	[0 1]	open+blind	EN, ES	SEALS

Open evaluation is made with already published reference alignments and blind evaluation is made by organizers, either from reference alignments unknown to the participants or manually.

3.1 Anatomy

The anatomy track comprises a single test case consisting of matching two fragments of biomedical ontologies which describe the human anatomy[5] (3304 classes) and the anatomy of the mouse[6] (2744 classes). The evaluation is based on a manually curated reference alignment. This dataset has been used since 2007 with some improvements over the years [13].

Systems are evaluated with the standard parameters of precision, recall, F-measure. Additionally, recall+ is computed by excluding trivial correspondences (i.e., correspondences that have the same normalized label). Alignments are also checked for coherence using the Pellet reasoner. The evaluation was carried out on a server with a 6 core CPU @ 3.46 GHz with 8GB allocated RAM, using the SEALS client. However, the evaluation parameters were computed *a posteriori*, after removing from the alignments produced by the systems s expressing relations other than equivalence, as well as trivial correspondences in the oboInOwl namespace (e.g., oboInOwl#Synonym = oboInOwl#Synonym). The results obtained with the SEALS client vary in some cases by 0.5% compared to the results presented below.

3.2 Biodiversity and Ecology

The new biodiversity track features two test cases based on highly overlapping ontologies that are particularly useful for biodiversity and ecology research: matching the Environment Ontology (ENVO) to the Semantic Web for Earth and Environment Technology Ontology (SWEET), and matching the Flora Phenotype Ontology (FLOPO) to the Plant Trait Ontology (PTO). The track was motivated by two projects, namely GFBio[7] (The German Federation for Biological Data) and AquaDiva[8], which aim at providing semantically enriched data management solutions for data capture, annotation, indexing and search [32]. Table 2 summarizes the versions and the sizes of the ontologies used in OAEI 2018.

Table 2. Versions and number of classes of the Biodiversity and Ecology track ontologies.

Ontology	Version	Classes
ENVO	2017-08-22	6909
SWEET	2018-03-12	4543
FLOPO	2016-06-03	24199
PTO	2017-09-11	1504

The reference alignments for the two test cases were produced through a hybrid approach that consisted of (1) using established matching systems to produce an au-

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

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

[7] www.gfbio.org

[8] www.aquadiva.uni-jena.de

tomated consensus alignment (akin to those used in the Disease and Phenotype track) then (2) manually validating the unique results produced by each system (and adding them to the consensus if deemed correct), and finally (3) adding manually generated correspondences. The matching systems used were the OAEI 2017 versions of AML, LogMap, LogMapBio, LogMapLite, LYAM, POMap, and YAMBio, in addition to the alignments from BioPortal [38].

The evaluation was carried out on a Windows 10 (64-bit) desktop with an Intel Core i5-7500 CPU @ 3.40GHz x 4 with 15.7 Gb RAM allocated, using the SEALS client. Systems were evaluated using the standard metrics.

3.3 Conference

The conference track features a single test case that is a suite of 21 matching tasks corresponding to the pairwise combination of 7 moderately expressive ontologies describing the domain of organizing conferences. The dataset and its usage are described in [47].

The track uses several reference alignments for evaluation: the old (and not fully complete) manually curated open reference alignment, *ra1*; an extended, also manually curated version of this alignment, *ra2*; a version of the latter corrected to resolve violations of conservativity, *rar2*; and an uncertain version of *ra1* produced through crowd-sourcing, where the score of each correspondences is the fraction of people in the evaluation group that agree with the correspondence. The latter reference was used in two evaluation modalities: *discrete* and *continuous* evaluation. In the former, correspondences in the uncertain reference alignment with a score of at least 0.5 are treated as correct whereas those with lower score are treated as incorrect, and standard evaluation parameters are used to evaluated systems. In the latter, weighted precision, recall and F-measure values are computed by taking into consideration the actual scores of the uncertain reference, as well as the scores generated by the matching system. For the sharp reference alignments (*ra1*, *ra2* and *rar2*), the evaluation is based on the standard parameters, as well the $F_{0.5}$-measure and F_2-measure and on conservativity and consistency violations. Whereas F_1 is the harmonic mean of precision and recall where both receive equal weight, F_2 gives higher weight to recall than precision and $F_{0.5}$ gives higher weight to precision higher than recall.

Two baseline matchers are use to benchmark the systems: edna string edit distance matcher; and StringEquiv string equivalence matcher as in the anatomy test case.

The evaluation was carried out on a Windows 10 (64-bit) desktop with an Intel Core i7–8550U (1,8 GHz, TB 4 GHz) x 4 with 16 GB RAM allocated using the SEALS client. Systems were evaluated using the standard metrics.

3.4 Disease and Phenotype

The Disease and Phenotype is organized by the Pistoia Alliance Ontologies Mapping project team[9]. It comprises 2 test cases that involve 4 biomedical ontologies covering the disease and phenotype domains: Human Phenotype Ontology (HP) versus

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

Mammalian Phenotype Ontology (MP) and Human Disease Ontology (DOID) versus Orphanet and Rare Diseases Ontology (ORDO). Currently, correspondences 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. More details about the Pistoia Alliance Ontologies Mapping project and the OAEI evaluation are available in [23]. Table 3.4 summarizes the versions of the ontologies used in OAEI 2018.

Table 3. 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

The reference alignments used in this track are silver standard consensus alignments automatically built by merging/voting the outputs of the participating systems in 2016, 2017 and 2018 (with vote=3). Note that systems participating with different variants and in different years only contributed once in the voting, that is, the voting was done by family of systems/variants rather than by individual systems. The HP-MP silver standard thus produced contains 2232 correspondences, whereas the DOID-ORDO one contains 2808 correspondences.

Systems were evaluated using the standard parameters as well as the number of unsatisfiable classes computed using the OWL 2 reasoner HermiT [36]. The evaluation was carried out in a Ubuntu 18 Laptop with an Intel Core i9-8950HK CPU @ 2.90GHz x 12 and allocating 25 Gb RAM.

3.5 Large Biomedical Ontologies

The large biomedical ontologies (largebio) track 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. The track consists of six test cases corresponding to three matching problems (FMA-NCI, FMA-SNOMED and SNOMED-NCI) in two modalities: small overlapping fragments and whole ontologies (FMA and NCI) or large fragments (SNOMED-CT).

The reference alignments used in this track are derived directly from the UMLS Metathesaurus [5] as detailed in [29], then automatically repaired to ensure logical coherence. However, rather than use a standard repair procedure of removing problem causing correspondences, we set the relation of such correspondences to *"?"* (unknown). These "?" correspondences are neither considered positive nor negative when evaluating matching systems, but are simply 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 avoid any bias,

correspondences were considered problem causing if they were selected for removal by any of the three established repair algorithms: Alcomo [34], LogMap [28], or AML [39]. The reference alignments are summarized in Table 4.

Table 4. 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

The evaluation was carried out in a Ubuntu 18 Laptop with an Intel Core i9-8950HK CPU @ 2.90GHz x 12 and allocating 25 Gb of RAM. Evaluation was based on the standard parameters (modified to account for the "?" relations) as well as the number of unsatisfiable classes and the ratio of unsatisfiable classes with respect to the size of the union of the input ontologies. Unsatisfiable classes were computed using the OWL 2 reasoner HermiT [36], or, in the cases in which HermiT could not cope with the input ontologies and the alignments (in less than 2 hours) a lower bound on the number of unsatisfiable classes (indicated by $\geq$) was computed using the OWL 2 EL reasoner ELK [33].

3.6 Multifarm

The multifarm track [35] aims at evaluating the ability of matching systems to deal with ontologies in different natural languages. This dataset results from the translation of 7 ontologies from the conference track (cmt, conference, confOf, iasted, sigkdd, ekaw and edas) into 10 languages: Arabic (ar), Chinese (cn), Czech (cz), Dutch (nl), French (fr), German (de), Italian (it), Portuguese (pt), Russian (ru), and Spanish (es). The dataset is composed of 55 pairs of languages, with 49 matching tasks for each of them, taking into account the alignment direction (e.g. $\text{cmt}_{en} \rightarrow \text{edas}_{de}$ and $\text{cmt}_{de} \rightarrow \text{edas}_{en}$ are distinct matching tasks). While part of the dataset is openly available, all matching tasks involving the *edas* and *ekaw* ontologies (resulting in 55×24 matching tasks) are used for blind evaluation.

We consider two test cases: i) those tasks where two different ontologies (cmt→edas, 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 only related to the use of specific techniques for dealing with cross-lingual ontologies, but also on the ability to exploit the identical structure of the ontologies.

The reference alignments used in this track derive directly from the manually curated Conference *ra1* reference alignments. Systems are evaluated using the standard parameters. The evaluation was carried out on a Ubuntu 16.04 machine configured with 16GB of RAM running under a i7-4790K CPU 4.00GHz x 8 processors, using the SEALS client.

3.7 IIMB

The new IIMB (ISLab Instance Matching Benchmark) track features a single test case consisting of 80 instance matching tasks, in which the goal is to match an original OWL Abox to an automatically transformed version of this Abox using the SWING (Semantic Web INstance Generation) framework [22]. SWING consists of a pool of transformation techniques organized as follows:

- *Data value transformations* (DVL) are based on changes of cardinality and content of property values belonging to instance descriptions (e.g. value deletion, value modification through random character insertion/substitution).
- *Data structure transformations* (DST) are based on changes of property names and structure within an instance description (e.g. string value splitting, property name modification).
- *Data semantics transformations* (DSS) are based on changes of class/type properties belonging to instance descriptions (e.g. property type deletion/modification).

The IIMB dataset has been generated by relying on a seed of linked-data instances I' extracted from the web. A set of manipulated instances I'' has been created from I' and inserted in IIMB by applying a combination of SWING transformation techniques according to the following schema:

- Tasks ID 001-020: DVL transformations
- Tasks ID 021-040: DST transformations
- Tasks ID 041-060: DSS transformations
- Tasks ID 061-080: DVL, DST, and DSS transformations

Within a group of tasks, the complexity of applied transformations increases with the task ID. In each task, the reference alignment corresponds to the correspondence-set generated by SWING between the instances of the original and transformed Abox.

The evaluation has been performed on an Intel Xeon E5/Core i7 server with 16GB RAM, the Ubuntu operating systems equipped with the SEALS client.

3.8 Link Discovery

The Link Discovery track features two test cases, Linking and Spatial, that deal with *link discovery* for spatial data represented as *trajectories* i.e., sequences of longitude, latitude pairs. The track is based on two datasets generated from TomTom[10] and Spaten [10].

The **Linking** test case aims at testing the performance of instance matching tools that implement mostly string-based approaches for identifying matching entities. It can be used not only by instance matching tools, but also by SPARQL engines that deal with query answering over geospatial data. The test case was based on SPIMBENCH [40], but since the ontologies used to represent trajectories are fairly simple and do not consider complex RDF or OWL schema constructs already supported by SPIMBENCH, only a subset of the transformations implemented by SPIMBENCH was used.

[10] `https://www.tomtom.com/en_gr/`

The transformations implemented in the test case were (I) string-based with different (a) levels, (b) types of spatial object representations and (c) types of date representations, and (II) schema-based, i.e., addition and deletion of ontology (schema) properties. These transformations were implemented in the TomTom dataset. In a nutshell, instance matching systems are expected to determine whether two traces with their points annotated with place names designate the same trajectory. In order to evaluate the systems we built a ground truth containing the set of expected links where an instance s_1 in the source dataset is associated with an instance t_1 in the target dataset that has been generated as a modified description of s_1.

The *Spatial* test case aims at testing the performance of systems that deal with topological relations proposed in the state of the art DE-9IM (Dimensionally Extended nine-Intersection Model) model [44]. The benchmark generator behind this test case implements all topological relations of DE-9IM between trajectories in the two dimensional space. To the best of our knowledge such a generic benchmark, that takes as input trajectories and checks the performance of linking systems for spatial data does not exist. For the design, we focused on (a) on the correct implementation of all the topological relations of the DE-9IM topological model and (b) on producing large datasets large enough to stress the systems under test. The supported relations are: *E*quals, Disjoint, Touches, Contains/Within, Covers/CoveredBy, Intersects, Crosses, Overlaps. The test case comprises tasks for all the DE-9IM relations and for LineString/LineString and LineString/Polygon cases, for both TomTom and Spaten datasets, ranging from 200 to 2K instances. We did not exceed 64 KB per instance due to a limitation of the Silk system[11], in order to enable a fair comparison of the systems participating in this track.

The evaluation for both test cases was carried out using the HOBBIT platform.

3.9 SPIMBENCH

The SPIMBENCH 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). The datasets were generated and transformed using SPIMBENCH [40] by altering a set of original linked data through value-based, structure-based, and semantics-aware transformations (simple combination of transformations). They share almost the same ontology (with some differences in property level, due to the structure-based transformations), which describes instances using 22 classes, 31 Data Properties, and 85 Object Properties. Participants are requested to produce a set of correspondences between the pairs of matching instances from the source and target datasets that are found to refer to the same real-world entity. An instance in the source dataset can have none or one matching counterparts in the target dataset. The SPIMBENCH task is composed of two datasets[12] with different scales (i.e., number of instances to match):

– Sandbox (380 INSTANCES, 10000 TRIPLES). It contains two datasets called source (Tbox1) and target (Tbox2) as well as the set of expected correspondences (i.e., reference alignment).

[11] `https://github.com/silk-framework/silk/issues/57`

[12] Although the files are called Tbox1 and Tbox2, they actually contain a Tbox and an Abox.

– Mainbox (1800 CWs, 50000 TRIPLES). It contains two datasets called source (Tbox1) and target (Tbox2). This test case is blind, meaning that the reference alignment is not given to the participants. In both datasets, the goal is to discover the correspondences among the instances in the source dataset (Tbox1) and the instances in the target dataset (Tbox2).

The evaluation was carried out using the HOBBIT platform.

3.10 Knowledge Graph

The new Knowledge Graph track consists of nine isolated graphs generated by running the DBpedia extraction framework on nine different Wikis from the Fandom Wiki hosting platform[13] in the course of the DBkWik project [24, 25]. These knowledge graphs cover three different topics, with three knowledge graphs per topic, so the track consists of nine test cases, corresponding to the pairwise combination of the knowledge graphs in each topic. The goal of each test case is to match both the instances and the schema simultaneously. The datasets are summarized in Table 5

Table 5. Characteristics of the Knowledge Graphs in the KG track, and the sources they were created from.

Source	Hub	Topic	#Instances	#Properties	#Classes
RuneScape Wiki	Games	Gaming	200,605	1,998	106
Old School RuneScape Wiki	Games	Gaming	38,563	488	53
DarkScape Wiki	Games	Gaming	19,623	686	65
Marvel Database	Comics	Comics	56,464	99	2
Hey Kids Comics Wiki	Comics	Entertainment	158,234	1,925	181
DC Database	Comics	Lifestyle	128,495	177	5
Memory Alpha	TV	Entertainment	63,240	326	0
Star Trek Expanded Universe	TV	Entertainment	17,659	201	3
Memory Beta	Books	Entertainment	63,223	413	11

The evaluation was based on a gold standard[14] of correspondences both on the schema and the instance level. While the schema level correspondences were created by experts, the instance correspondences were crowd sourced using Amazon MTurk. Since we do not have a correspondence for each instance, class, and property in the graphs, this gold standard is only a *partial gold standard*.

The evaluation was executed on a virtual machine (VM) with 32GB of RAM and 16 vCPUs (2.4 GHz), with Debian 9 operating system and Openjdk version 1.8.0_181, using the SEALS client. It was not executed on the HOBBIT platform because few systems registered in HOBBIT for this task and all of them also had a SEALS counterpart.

We used the `-o` option in SEALS (version 7.0.5) to provide the two knowledge graphs which should be matched. We used local files rather than HTTP URLs to circumvent the overhead of downloading the knowledge graphs. We could not use the

[13] `https://www.wikia.com/`

[14] `http://dbkwik.webdatacommons.org`

"-x" option of SEALS because we had to modify the evaluation routine for two reasons. First, we wanted to differentiate between results for class, property, and instance correspondences, and second, we had to change the evaluation to deal with the partial nature of our gold standard.

The alignments were evaluated based on precision, recall, and f-measure for classes, properties, and instances (each in isolation). Our partial gold standard contained 1:1 correspondences, as well as *negative* correspondences, i.e., correspondences stating that a resource `A` in one knowledge graph has *no* correspondence in the second knowledge graph. This allows to increase the count of false positives if the matcher nevertheless finds a correspondence (i.e., maps `A` to a resource in the other knowledge graph). We further assume that in each knowledge graph, only one representation of the concept exists. This means that if we have a correspondence in our gold standard, we count a correspondence to a different concept as a false positive. The count of false negatives is only increased if we have a 1:1 correspondence and it is not found by a matcher. The whole source code for generating the evaluation results is also available[15].

As a benchmark, we employed a simple string matching approach with some out of the box text preprocessing to generate a baseline. The source code for this approach is publicly available[16].

3.11 Interactive Matching

The interactive matching track aims to assess the performance of semi-automated matching systems by simulating user interaction [37, 12]. The evaluation thus focuses on how interaction with the user improves the matching results. Currently, this track does not evaluate the user experience or the user interfaces of the systems [26, 12].

The interactive matching track is based on the datasets from the Anatomy and Conference tracks, which have been previously described. It relies on the SEALS client's *Oracle* class to simulate user interactions. An interactive matching system can present a collection of correspondences simultaneously to the oracle, which will tell the system whether that correspondence is correct or not. If a system presents up to three correspondences together and each correspondence presented has a mapped entity (i.e., class or property) in common with at least one other correspondence 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 correspondences. 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.

In addition to the standard evaluation parameters, we also compute the number of requests made by the system, the total number of distinct correspondences asked, the number of positive and negative answers from the oracle, the performance of the system according to the oracle (to assess the impact of the oracle errors on the system) and finally, the performance of the oracle itself (to assess how erroneous it was).

[15] `http://oaei.ontologymatching.org/2018/results/knowledgegraph/kg_track_eval.zip`

[16] `http://oaei.ontologymatching.org/2018/results/knowledgegraph/string_baseline_kg-source.zip`

The evaluation was carried out 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 is the total number of interactions for all the pairs.

3.12 Complex Matching

The complex matching track is meant to evaluate the matchers based on their ability to generate complex alignments. A complex alignment is composed of complex correspondences typically involving more than two ontology entities, such as o_1*:AcceptedPaper* $\equiv$ o_2*:Paper* $\sqcap$ o_2*:hasDecision.*o_2*:Acceptance*. Four datasets with their own evaluation process have been proposed [46].

The **complex conference** dataset is composed of three ontologies: cmt, conference and ekaw from the conference dataset. The reference alignment was created as a consensus between experts. In the evaluation process, the matchers can take the simple reference alignment *ra1* as input. The precision and recall measures are manually calculated over the complex equivalence correspondences only.

The **Hydrography** dataset consists of matching four different source ontologies (hydro3, hydrOntology-translated, hydrOntology-native, and cree) to a single target ontology (SWO). The evaluation process is based on three subtasks: given an entity from the source ontology, identify all related entities in the source and target ontology; given an entity in the source ontology and the set of related entities, identify the logical relation that holds between them; identify the full complex correspondences. The first subtask was evaluated based on precision and recall and the latter two were evaluated using semantic precision and recall.

The **GeoLink** dataset derives from the homonymous project, funded under the U.S. National Science Foundation's EarthCube initiative. It is composed of two ontologies: the GeoLink Base Ontology (GBO) and the GeoLink Modular Ontology (GMO). The GeoLink project is a real-world use case of ontologies, and instance data is available. The alignment between the two ontologies was developed in consultation with domain experts from several geoscience research institutions. More detailed information on this benchmark can be found in [48]. Evaluation was done in the same way as with the Hydrography dataset. The evaluation platform was a MacBook Pro with a 2.6 GHz Intel Core i5 processor and 16 GB of 1600 MHz DDR3 RAM running macOS Mojave version 10.14.2.

The **Taxon** dataset is composed of four knowledge bases containing knowledge about plant taxonomy: AgronomicTaxon, AGROVOC, TAXREF-LD and DBpedia. The evaluation is two-fold: first, the precision of the output alignment is manually assessed; then, a set of source queries are rewritten using the output alignment. The rewritten target query is then manually classified as correct or incorrect. A source query is considered successfully rewritten if at least one of the target queries is semantically equivalent to it. The proportion of source queries successfully rewritten is then calculated (QWR in the results table). The evaluation over this dataset is open to all matching systems

(simple or complex) but some queries can not be rewritten without complex correspondences. The evaluation was performed with an Ubuntu 16.04 machine configured with 16GB of RAM running under a i7-4790K CPU 4.00GHz x 8 processors.

4 Results and Discussion

4.1 Participation

Following an initial period of growth, the number of OAEI participants has remained approximately constant since 2012, at slightly over 20. This year we observed a slight decrease to 19 participating systems. Table 6 lists the participants and the tracks in which they competed. Some matching systems participated with different variants (AML, LogMap) whereas others were evaluated with different configurations, as requested by developers (see test case sections for details).

Table 6. Participants and the status of their submissions.

System	ALIN	ALOD2vec	AML	AMLC	CANARD	DOME	EVOCROS	FCAMapX	Holontology	KEPLER	Lily	LogMap	LogMap-Bio	LogMapLt	POMAP++	RADON	SANOM	Silk	XMap	Total=19
Confidence	-	-	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	14
Anatomy	●	●	●	○	○	●	○	●	●	●	●	●	●	●	●	○	●	○	●	14
Biodiversity & Ecology	○	○	●	○	○	○	○	○	○	◐	●	●	●	●	●	○	○	○	●	8
Conference	●	●	●	○	○	●	○	●	●	●	●	●	○	●	○	○	●	○	●	12
Disease & Phenotype	○	○	●	○	○	●	○	○	○	◐	●	●	●	●	●	○	○	○	●	9
Large Biomedical Ont.	○	◐	●	○	○	●	○	●	○	◐	○	●	●	●	◐	○	○	○	●	10
Multifarm	○	○	●	○	○	◐	◐	○	○	◐	○	●	○	○	○	○	○	○	●	6
IIMB	○	○	●	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	2
Link Discovery	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	◐	○	◐	○	3
SPIMBENCH	○	○	●	○	○	○	○	○	○	○	●	●	○	○	○	○	○	○	○	3
Knowledge Graph	○	○	◐	○	○	●	○	○	●	○	○	◐	●	●	●	○	○	○	○	7
Interactive Matching	●	○	●	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	●	4
Complex Matching	○	◐	◐	◐	◐	◐	○	◐	◐	◐	○	◐	◐	◐	◐	○	○	○	◐	13
total	3	4	12	1	1	7	1	4	4	7	5	11	6	7	6	1	2	1	8	65

Confidence pertains to the confidence 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.

A number of participating systems use external sources of background knowledge, which are especially critical in matching ontologies in the biomedical domain. LogMap-Bio uses BioPortal as mediating ontology provider, that is, it retrieves from BioPortal

the most suitable top-10 ontologies for each matching task. LogMap uses normalizations and spelling variants from the general (biomedical) purpose SPECIALIST Lexicon. 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). XMAP and Lily use a dictionary of synonyms (pre)extracted from the UMLS Metathesaurus. In addition Lily also uses a dictionary of synonyms (pre)extracted from BioPortal.

4.2 Anatomy

The results for the Anatomy track are shown in Table 7.

Table 7. Anatomy results, ordered by F-measure. Runtime is measured in seconds; "size" is the number of correspondences in the generated alignment.

System	Runtime	Size	Precision	F-measure	Recall	Recall+	Coherent
AML	42	1493	0.95	0.943	0.936	0.832	√
LogMapBio	808	1550	0.888	0.898	0.908	0.756	√
POMAP++	210	1446	0.919	0.897	0.877	0.695	-
XMap	37	1413	0.929	0.896	0.865	0.647	√
LogMap	23	1397	0.918	0.88	0.846	0.593	√
SANOM	487	1450	0.888	0.865	0.844	0.632	-
FCAMapX	118	1274	0.941	0.859	0.791	0.455	-
KEPLER	244	1173	0.958	0.836	0.741	0.316	-
Lily	278	1382	0.872	0.832	0.795	0.518	-
LogMapLite	18	1147	0.962	0.828	0.728	0.288	-
ALOD2Vec	75	987	0.996	0.785	0.648	0.086	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
DOME	22	935	0.997	0.761	0.615	0.009	-
ALIN	271	928	0.998	0.758	0.611	0.0	√
Holontology	265	456	0.976	0.451	0.294	0.005	-

Of the 14 systems participating in the Anatomy track, 11 achieved an F-measure higher than the StringEquiv baseline. Three systems were first time participants (ALOD2Vec, DOME, and Holontology) and showed modest results in terms of both F-measure and recall+, with only ALOD2Vec ranking above the baseline. Among the five systems that participated for the second time, SANOM shows increases in both F-measure (from 0.828 to 0.865) and recall+ (from 0.419 to 0.632), KEPLER and Lily have the same performance as last year, and both POMAP++ (POMap in 2017) and FCAMapX (FCA_Map in 2016) have decreases in F-measure and recall+. Long-term systems showed few changes in comparison with previous years with respect to alignment quality (precision, recall, F-measure, and recall+), size or run time. The exceptions were LogMapBio which increased in both recall+ (from 0.733 to 0.756) and alignment size (by 16 correspondences) since last year, and ALIN that had a substantial increase of 412 correspondences since last year.

In terms of run time, 6 out of 14 systems computed an alignment in less than 100 seconds, a ratio which is similar to 2017 (5 out of 11). LogMapLite remains the system with the shortest runtime. Regarding quality, AML remains the system with the highest F-measure (0.943) and recall+ (0.832), but 4 other systems obtained an F-measure above 0.88 (LogMapBio, POMap++, XMap, and LogMap) which is at least as good as the best systems in OAEI 2007-2010. Like in previous years, there is no significant correlation between the quality of the generated alignment and the run time. Five systems produced coherent alignments, which is the same as last year.

4.3 Biodiversity and Ecology

Of the 8 participants registered for this track, 7 systems (AML, LogMap, LogMapBio, LogMapLt, Lily, XMap and POMap) managed to generate a meaningful output in 4 hours, and only KEPLER did not. Table 8 shows the results for the FLOPO-PTO and ENVO-SWEET tasks.

Table 8. Results for the Biodiversity & Ecology track, ordered by F-measure.

System	Size	Precision	F-measure	Recall
FLOPO-PTO task				
AML	233	0.88	0.86	0.84
LogMap	235	0.817	0.802	0.787
LogMapBio	239	0.803	0.795	0.787
XMap	153	0.987	0.761	0.619
LogMapLite	151	0.987	0.755	0.611
POMap	261	0.663	0.685	0.709
LiLy	176	0.813	0.681	0.586
ENVO-SWEET task				
AML	791	0,776	0,844	0,926
LogMap	583	0,839	0,785	0,738
POMap	583	0,839	0,785	0,738
XMap	547	0,868	0,785	0,716
LogMapBio	572	0,839	0,777	0,724
LogMapLite	740	0,732	0,772	0,817
LiLy	491	0,866	0,737	0,641

Regarding the FLOPO-PTO task, the top 3 ranked systems in terms of F-measure are AML, LogMap and LogMapBio, with curiously a similar number of generated correspondences among them. Among these, AML achieved the highest F-measure (0.86) and a well-balanced result, with over 80% recall and a still quite high precision.

Regarding the ENVO-SWEET task, AML ranked first in terms of F-measure, followed by a three-way tie between LogMap, POMAP and XMap. AML had a less balanced result in this test case, with a very high recall and significant larger alignment than the other top systems, but a comparably lower precision. LogMap and POMap

produced alignments of exactly equal size and quality, whereas XMap had the highest precision overall, but a lower recall than these.

Overall, in this first evaluation, the results obtained from participating systems are quite promising, as all systems achieved more than 0.68 in term of F-measure. We should note that most of the participating systems, and all of the most successful ones use external resources as background knowledge.

4.4 Conference

The conference evaluation results using the sharp reference alignment *rar2* are shown in Table 9. For the sake of brevity, only results with this reference alignment and considering both classes and properties are shown. For more detailed evaluation results, please check conference track's web page.

With regard to the two baselines we can group the twelve participants into four groups: six matching systems outperformed both baselines (SANOM, AML, LogMap, XMap, FCAMapX and DOME); three performed the same as the edna baseline (ALIN, LogMapLt and Holontology); two performed slightly worse than this baseline (KEPLER and ALOD2Vec); and Lily performed worse than both baselines. Note that two systems (ALIN and Lily) do not match properties at all which naturally has a negative effect on their overall performance.

The performance of all matching systems regarding their precision, recall and F_1-measure is plotted in Figure 1. Systems are represented as squares or triangles, whereas the baselines are represented as circles.

With respect to logical coherence [42, 43], only three tools (ALIN, AML and LogMap) have no consistency principle violation (in comparison to five tools last year and seven tools two years ago). This year all tools have some conservativity principle violations (in comparison to one tool having no conservativity principle violation last year). 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.

The Conference evaluation results using the uncertain reference alignments are presented in Table 10.

Among the twelve participating alignment systems, six use 1.0 as the confidence value for all matches they identify (ALIN, ALOD2Vec, DOME, FCAMapX, Holontology, LogMapLt), whereas the remaining six have a wide range of confidence values (AML, KEPLER, Lily, LogMap, SANOM and XMap).

When comparing the performance of the matchers on the uncertain reference alignments versus that on the sharp version (with the corresponding *ra1*), we see that in the discrete case all matchers except Lily performed the same or better in terms of F-measure (Lily's F-measure dropped by 0.01). The changes in F-measure ranged from -1 to 15 percent over the sharp reference alignment. This was predominantly 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 the matchers with confidence values always 1.0 is very similar regardless of whether a discrete or continuous evaluation methodology is used, because many of their correspondences are ones that the experts had high agreement

Table 9. 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.

System	Prec.	$F_{0.5}$-m.	F_1-m.	F_2-m.	Rec.	Inc.Align.	Conser.V.	Consist.V.
SANOM	0.72	0.71	0.7	0.69	0.68	9	103	92
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	0.76	0.7	0.62	0.56	0.52	4	53	14
FCAMapX	0.64	0.62	0.59	0.56	0.54	11	124	273
DOME	0.74	0.66	0.57	0.5	0.46	3	106	10
edna	*0.74*	*0.66*	*0.56*	*0.49*	*0.45*			
ALIN	0.82	0.69	0.56	0.48	0.43	0	2	0
Holontology	0.73	0.65	0.56	0.49	0.45	3	66	10
LogMapLt	0.68	0.62	0.56	0.5	0.47	5	96	25
ALOD2Vec	0.67	0.62	0.55	0.5	0.47	6	124	27
KEPLER	0.67	0.61	0.55	0.49	0.46	12	123	159
StringEquiv	*0.76*	*0.65*	*0.53*	*0.45*	*0.41*			
Lily	0.54	0.53	0.52	0.51	0.5	9	140	124

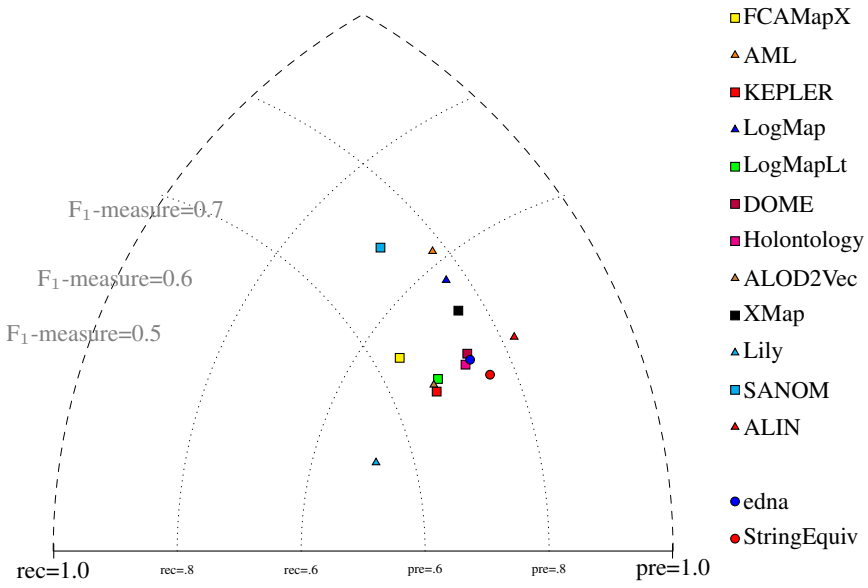

Fig. 1. Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F_1-measure are depicted by areas bordered by corresponding lines F_1-measure=0.$[5|6|7]$.

Table 10. F-measure, precision, and recall of matchers when evaluated using the sharp (*ra1*), discrete uncertain and continuous uncertain metrics. Sorted according to F_1-m. in continuous.

System	Sharp			Discrete			Continuous		
	Prec.	F_1-m.	Rec.	Prec.	F_1-m.	Rec.	Prec.	F_1-m.	Rec.
AML	0.84	0.74	0.66	0.79	0.78	0.77	0.80	0.77	0.74
SANOM	0.79	0.74	0.69	0.71	0.74	0.78	0.65	0.72	0.81
ALIN	0.88	0.60	0.46	0.88	0.69	0.57	0.88	0.70	0.59
XMap	0.81	0.65	0.54	0.66	0.74	0.83	0.74	0.70	0.66
DOME	0.79	0.60	0.48	0.79	0.68	0.60	0.78	0.69	0.62
Holontology	0.78	0.59	0.48	0.78	0.68	0.60	0.78	0.68	0.61
ALOD2Vec	0.71	0.59	0.50	0.71	0.66	0.62	0.71	0.67	0.63
LogMap	0.82	0.69	0.59	0.77	0.73	0.70	0.80	0.67	0.57
LogMapLt	0.73	0.59	0.50	0.73	0.67	0.62	0.72	0.67	0.63
FCAMapX	0.68	0.61	0.56	0.65	0.66	0.67	0.64	0.66	0.68
KEPLER	0.76	0.59	0.48	0.76	0.67	0.60	0.58	0.63	0.68
Lily	0.59	0.56	0.53	0.52	0.55	0.59	0.59	0.32	0.22

about, while the ones they missed were more controversial. AML produces a fairly wide range of confidence values and has the highest F-measure under both the continuous and discrete evaluation methodologies, indicating that this system's confidence evaluation does a good job of reflecting cohesion among experts on this task. Of the remaining systems, four (KEPLER, LogMap, SANOM and XMap) have relatively small drops in F-measure when moving from discrete to continuous evaluation. Lily's performance drops drastically under the continuous evaluation methodology. This is because the matcher assigns low confidence values to some correspondences in which the labels are equivalent strings, which many experts agreed with unless there was a compelling reason not to. This hurts recall, but using a low threshold value in the discrete version of the evaluation metrics 'hides' this problem.

Overall, in comparison with last year, the F-measures of most returning matching systems essentially held constant under both the sharp and uncertain evaluations. The exceptions were ALIN and SANOM, whose performance improved substantially. In fact, the latter improved its performance so much that it became the top system with regard to F-measure according to the sharp evaluation. We can conclude that all matchers perform better on the fuzzy versus sharp version of the benchmark and that the performance of AML against the fuzzy reference alignment rivals that of a human evaluated in the same way.

4.5 Disease and Phenotype Track

In the OAEI 2018 phenotype track 9 systems were able to complete at least one of the tasks with a 6 hours timeout. Tables 11 show the evaluation results in the HP-MP and DOID-ORDO matching tasks, respectively.

Since the consensus reference alignments only allow us to assess how systems perform in comparison with one another, the proposed ranking is only a reference. Note that some of the correspondences in the consensus alignment may be erroneous (false

Table 11. Results for the HP-MP and DOID-ORDO tasks based on the consensus reference alignment.

System	Time (s)	# Corresp.	# Unique	Scores			Incoherence	
				Prec.	F-m.	Rec.	Unsat.	Degree
HP-MP task								
LogMap	31	2,130	1	0.88	0.86	0.84	0	0.0%
LogMapBio	821	2,178	37	0.86	0.85	0.84	0	0.0%
AML	70	2,010	279	0.89	0.84	0.80	0	0.0%
LogMapLt	7	1,370	3	0.99	0.76	0.61	0	0.0%
POMAP++	1,668	1,502	214	0.86	0.69	0.58	0	0.0%
Lily	4,749	2,118	733	0.68	0.66	0.65	0	0.0%
XMap	20	704	2	0.99	0.48	0.31	0	0.0%
DOME	46	689	0	1.00	0.47	0.31	0	0.0%
DOID-ORDO task								
LogMap	25	2,323	0	0.94	0.85	0.78	0	0.0%
LogMapBio	1,891	2,499	91	0.90	0.85	0.80	0	0.0%
POMAP++	2,264	2,563	174	0.87	0.83	0.80	0	0.0%
LogMapLt	7	1,747	16	0.99	0.76	0.62	0	0.0%
XMap	15	1,587	37	0.97	0.70	0.55	0	0.0%
KEPLER	2,746	1,824	158	0.88	0.70	0.57	0	0.0%
Lily	2,847	3,738	1,167	0.59	0.67	0.78	206	1.9%
AML	135	4,749	1,886	0.51	0.65	0.87	0	0.0%
DOME	10	1,232	2	1.00	0.61	0.44	0	0.0%

positives) because all systems that agreed on it could be wrong (e.g., in erroneous correspondences with equivalent labels, which are not that uncommon in biomedical tasks). In addition, the consensus alignments will not be complete, because there are likely to be correct correspondences that no system is able to find, and there are a number of correspondences found by only one system (and therefore not in the consensus alignments) which may be correct. Nevertheless, the results with respect to the consensus alignments do provide some insights into the performance of the systems.

Overall, LogMap is the system that provides the closest set of correspondences to the consensus (not necessarily the best system) in both tasks. It has a small set of unique correspondences as most of its correspondences are also suggested by its variant LogMapBio and vice versa. By contrast, Lily and AML produce the highest number of unique correspondences in HP-MP and DOID-ORDO respectively, and the second-highest inversely. All systems produce coherent alignments except for Lily in the DOID-ORDO task.

4.6 Large Biomedical Ontologies

In the OAEI 2018 Large Biomedical Ontologies track, 10 systems were able to complete at least one of the tasks within a 6 hours timeout. Seven systems were able to complete all six tasks.[17] Since the reference alignments for this track are based on the

[17] Check out the supporting scripts to reproduce the evaluation: `https://github.com/ernestojimenezruiz/oaei-evaluation`

Table 12. Results for the whole ontologies matching tasks in the OAEI largebio track.

System	Time (s)	# Corresp.	# Unique	Scores			Incoherence	
				Prec.	F-m.	Rec.	Unsat.	Degree
Whole FMA and NCI ontologies (Task 2)								
AML	55	2,968	311	0.84	0.86	0.87	2	0.014%
LogMap	1,072	2,701	0	0.86	0.83	0.81	2	0.014%
LogMapBio	1,072	2,860	39	0.83	0.83	0.83	2	0.014%
XMap2	65	2,415	52	0.88	0.80	0.74	2	0.014%
FCAMapX	881	3,607	443	0.67	0.74	0.84	8,902	61.8%
LogMapLt	6	3,458	250	0.68	0.74	0.82	5,170	35.9%
DOME	12	2,383	10	0.80	0.73	0.67	596	4.1%
Whole FMA ontology with SNOMED large fragment (Task 4)								
FCAMapX	1,736	7,971	1,258	0.82	0.79	0.76	21,289	57.0%
AML	94	6,571	462	0.88	0.77	0.69	0	0.0%
LogMapBio	1,840	6,471	31	0.83	0.73	0.65	0	0.0%
LogMap	288	6,393	0	0.84	0.73	0.65	0	0.0%
XMap2	299	6,749	1,217	0.72	0.66	0.61	0	0.0%
LogMapLt	9	1,820	56	0.85	0.33	0.21	981	2.6%
DOME	20	1,588	1	0.94	0.33	0.20	951	2.5%
Whole NCI ontology with SNOMED large fragment (Task 6)								
AML	168	13,176	1,230	0.90	0.77	0.67	$\geq$517	$\geq$0.6%
FCAMapX	2,377	15,383	1,670	0.80	0.73	0.68	$\geq$72,859	$\geq$85.5%
LogMapBio	2,942	13,098	231	0.85	0.72	0.63	$\geq$3	$\geq$0.004%
LogMap	475	12,276	0	0.87	0.71	0.60	$\geq$1	$\geq$0.001%
LogMapLt	11	12,864	720	0.80	0.66	0.57	$\geq$74,013	$\geq$86.9%
DOME	24	9,702	42	0.91	0.63	0.49	$\geq$53,574	$\geq$62.9%
XMap2	427	16,271	4,432	0.64	0.61	0.58	$\geq$73,571	$\geq$86.4%

UMLS-Metathesaurus, we disallowed the use of this resource as a source of background knowledge in the matching systems that used it, XMap and Lily. XMap was still able to produce competitive results, while Lily produced an empty set of alignments. The evaluation results for the largest matching tasks are shown in Tables 12.

The top-ranked systems by F-measure were respectively: AML and LogMap in Task 2; FCAMapX and AML in Task 4; and AML and FCAMapX in Task 6.

Interestingly, the use of background knowledge led to an improvement in recall from LogMap-Bio over LogMap in all tasks, but this came at the cost of precision, resulting in the two variants of the system having very similar F-measures.

The effectiveness of all systems decreased from small fragments to whole ontologies tasks.[18] One reason for this is that with larger ontologies there are more plausible correspondence candidates, and thus it is harder to attain both a high precision and a high recall. In fact, this same pattern is observed moving from the FMA-NCI to the FMA-SNOMED to the SNOMED-NCI problem, as the size of the task also increases. Another reason is that the very scale of the problem constrains the matching strategies

[18] http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2018/results/

that systems can employ: AML for example, forgoes its matching algorithms that are computationally more complex when handling very large ontologies, due to efficiency concerns.

The size of the whole ontologies tasks proved a problem for a number of systems, which were unable to complete them within the allotted time: POMAP++, ALOD2Vec and KEPLER.

With respect to alignment coherence, as in previous OAEI editions, 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 [34]). 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). XMap seems to deactivate the repair facility for the SNOMED-NCI case.

As the results tables show, 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 [34], the repair module of LogMap (LogMap-Repair) [28] or the repair module of AML [39], which have worked well in practice [30, 21].

4.7 Multifarm

This year, 6 matching systems registered for the MultiFarm track: AML, DOME, EVOCROS, KEPLER, LogMap and XMap. This represents a slight decrease from the last two years, but is within an approximately constant trend (8 in 2017, 7 in 2016, 5 in 2015, 3 in 2014, 7 in 2013, and 7 in 2012). However, a few systems had issues when evaluated: i) KEPLER generated some parsing errors for some pairs; ii) EVOCROS took around 30 minutes to complete a single task (we have hence tested only 50 matching tasks) and generated empty alignments; iii) DOME was not able to generate any alignment; iv) XMap had problems dealing with most pairs involving the ar, ru and cn languages. Please refer to the OAEI papers of the matching systems for a detailed description of the strategies employed by each system, most of which adopt a translation step before the matching itself.

The Multifarm evaluation results based on the blind dataset are presented in Table 13. They have been computed using the Alignment API 4.9 and can slightly differ from those computed with the SEALS client. We do not report the results of non-specific systems here, as we could observe in the last campaigns that they can have intermediate results in the "same ontologies" task (ii) and poor performance in the "different ontologies" task (i).

With respect to run time, we observe large differences between systems due to the high number of matching tasks involved (55 x 24). Note as well that the concurrent access to the SEALS repositories during the evaluation period may have an impact on the time required for completing the tasks.

Table 13. 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	26	55	6.87	.72 (.72)	.46 (.46)	.35 (.35)	23.24	.96 (.95)	.27 (.27)	.16 (.16)
KEPLER	900	53	9.74	.40 (.42)	.27 (.28)	.21 (.22)	58.28	.85 (.88)	.49 (.51)	.36 (.37)
LogMap	39	55	6.99	.72 (.72)	.37 (.37)	.25 (.25)	46.80	.95 (.96)	.41 (.42)	.28 (.28)
XMap	22	26	94.72	.02 (.05)	.03 (.07)	.07 (.07)	345.00	.13 (.18)	.14 (.20)	.19 (.19)

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 generate (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 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.

In terms of F-measure, AML remains the top performing system in task (i), followed by LogMap and KEPLER. In task (ii), AML has relatively low performance (with a notably low recall) and KEPLER has the highest F-measure, followed by LogMap.

With respect to the pairs of languages for test cases of type (i), for the sake of brevity, we do not present the detailed results. Please refer to the OAEI results web page to view them. The language pairs in which systems perform better in terms of F-measure include: es-it, it-pt and nl-pt (AML); cz-pt and de-pt (KEPLER); en-nl (LogMap); and cz-en (XMap). We note also some patterns behind the worst results obtained by systems: ar-cn for AML, and some pairs involving cn for KEPLER and LogMap)

In terms of performance, the F-measure for blind tests remains relatively stable across campaigns. AML and LogMap keep their positions and have similar F-measure with respect to the previous campaigns, as does XMap. As observed in previous campaigns, systems privilege precision over recall, and the results are expectedly below the ones obtained for the original Conference dataset. Cross-lingual approaches remain mainly based on translation strategies and the combination of other resources (like cross-lingual links in Wikipedia, BabelNet, etc.) while strategies such as machine learning, or indirect alignment composition remain under-exploited.

4.8 IIMB

Only two systems participated in the new IIMB track: AML and LogMap. The obtained results are summarized in Table 14[19].

In the results of both AML and LogMap, we note that high-quality performances are provided on test-cases based on DVL transformations. We note that the evaluation results on this kind of matching issues have been improved in the recent years (for instance, see [3] for a comparison against the 2012 version of the IIMB dataset). As a matter of fact, recognition of similarities across instance descriptions with data-value

[19] A detailed report of test-case results is provided on `https://islab.di.unimi.it/im_oaei_2018/`.

Table 14. Summary of the IIMB results.

System	Runtime (s)	Precision	Recall	F-measure
Data Value Transformations				
AML	1828	0.893	0.789	0.828
LogMap	4.2	0.896	0.893	0.889
Data Structure Transformations				
AML	2036	0.419	0.433	0.424
LogMap	5.7	0.934	0.985	0.959
Data Semantics Transformations				
AML	6.2	0.747	0.889	0.796
LogMap	4.6	0.855	0.947	0.893
Mixed Transformations				
AML	2083	0.334	0.294	0.295
LogMap	6.5	0.920	0.758	0.819

heterogeneities represents a sort of consolidated matching capability that can be considered as a standard functionality of the current state-of-the-art tools. We also note that promising results are also provided by both the participating tools on test-cases based on DSS transformations. We argue that such a kind of result is due to the capability of both AML and LogMap to cope with incoherence, thus reducing the number of false-positive results. As a final remark, we observe that recall is usually lower than precision. Maybe, the cause is the non-uniform quality of expected automatically-generated correspondences. Expected correspondences are created by applying a sequence of transformations with different length (i.e., number of transformations) and different degree of complexity (i.e., strength of applied data manipulations). Sometimes, the applied SWING transformations produce correspondences that are more difficult to agree with, rather than to detect. Measuring the quality of automatically-generated alignments as well as pruning of excessively-hard ones from the set of expected results is a challenging issue to consider in future research work (see Section 5).

4.9 Link Discovery

This year the Link Discovery track counted one participant in the Linking test case (AML) and three participants in the Spatial test case: AML, Silk and RADON.

In the Linking test case, AML perfectly captures all the correct links while not producing wrong ones, thus obtaining perfect precision and a recall (1.0) in both the Sandbox and Mainbox datasets. It required 6.8s and 313s, respectively, to complete the two tasks.

We divided the Spatial test cases into four suites. In the first two suites (SLL and LLL), the systems were asked to match LineStrings to LineStrings considering a given relation for 200 and 2K instances for the TomTom and Spaten datasets. In the last two tasks (SLP, LLP), the systems were asked to match LineStrings to Polygons (or Polygons to LineStrings depending on the relation) again for both datasets. Since the precision, recall and f-measure results from all systems were equal to 1.0, we are only

presenting results regarding the time performance. The time performance of the matching systems in the SLL, LLL, SLP and LLP suites are shown in Figures 2-3.

In the SLL suite, RADON has the best performance in most cases except for the *Touches* and *Intersects* relations, followed by AML. Silk seems to need the most time, particularly for *Touches* and *Intersects* relations in the TomTom dataset and *Overlaps* in both datasets.

In the LLL suite we have a more clear view of the capabilities of the systems with the increase in the number of instances. In this case, RADON and Silk have similar behavior as in the the small dataset, but it is more clear that the systems need much more time to match instances from the TomTom dataset. RADON has still the best performance in most cases. AML has the next best performance and is able to handle some cases better than other systems (e.g. *Touches* and *Intersects*), however, it also hits the platform time limit in the case of *Disjoint*.

In the SLP suite, in contrast to the first two suites, RADON has the best performance for all relations. AML and Silk have minor time differences and, depending on the case, one is slightly better than the other. All the systems need more time for the TomTom dataset but due to the small size of the instances the time difference is minor.

In the LLP suite, RADON again has the best performance in all cases. AML hits the platform time limit in *Disjoint* relations on both datasets and is better than Silk in most cases except *Contains* and *Within* on the TomTom dataset where it needs an excessive amount of time.

Taking into account the executed test cases we can identify the capabilities of the tested systems as well as suggest some improvements. All the systems participated in most of the test cases, with the exception of Silk which did not participate in the *Covers* and *Covered By* test cases.

RADON was the only system that successfully addressed all the tasks, and had the best performance for the SLP and LLP suites, but it can be improved for the *Touches* and *Intersects* relations for the SLL and LLL suites. AML performs extremely well in most cases, but can be improved in the cases of *Covers/Covered By* and *Contains/Within* when it comes to LineStrings/Polygons Tasks and especially in *Disjoint* relations where it hits the platform time limit. Silk can be improved for the *Touches*, *Intersects* and *Overlaps* relations and for the SLL and LLL tasks and for the *Disjoint* relation in SLP and LLP Tasks.

In general, all systems needed more time to match the TomTom dataset than the Spaten one, due to the smaller number of points per instance in the latter. Comparing the LineString/LineString to the LineString/Polygon Tasks we can say that all the systems needed less time for the first for the *Contains*, *Within*, *Covers* and *Covered by* relations, more time for the *Touches*, *Instersects* and *Crosses* relations, and approximately the same time for the *Disjoint* relation.

4.10 SPIMBENCH

This year, the SPIMBENCH track counted three participants: AML, Lily, and LogMap. The evaluation results of the track are shown in Table 15.

Lily had the best performance overall both in terms of F-measure and in terms of run time. Notably, its run time scaled very well with the increase in the number of

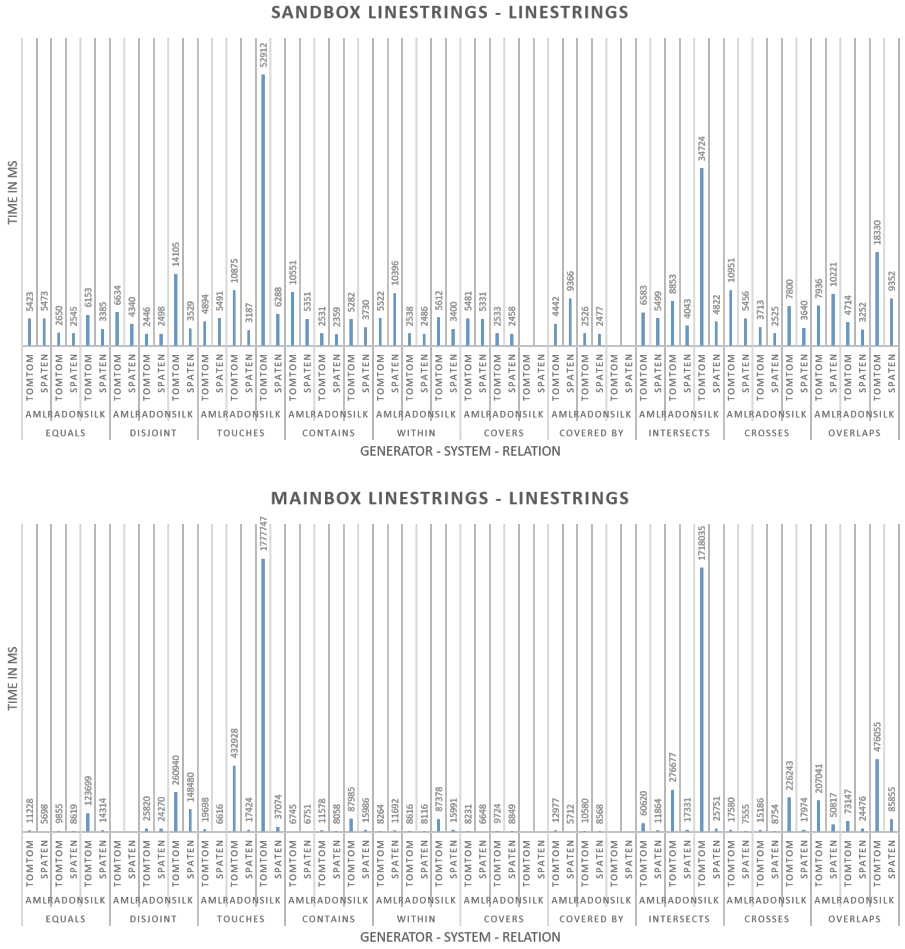

Fig. 2. Time performance for TomTom & Spaten SLL (top) and LLL (bottom) suites for AML (A), Silk (S) and RADON (R).

instances. Both Lily and AML had a higher recall than precision, with the former having full recall. By contrast, LogMap had the highest precision but lowest recall of the three systems. AML and LogMap had a similar run time for the Sandbox task, but the latter scaled better with the increase in the number of instances.

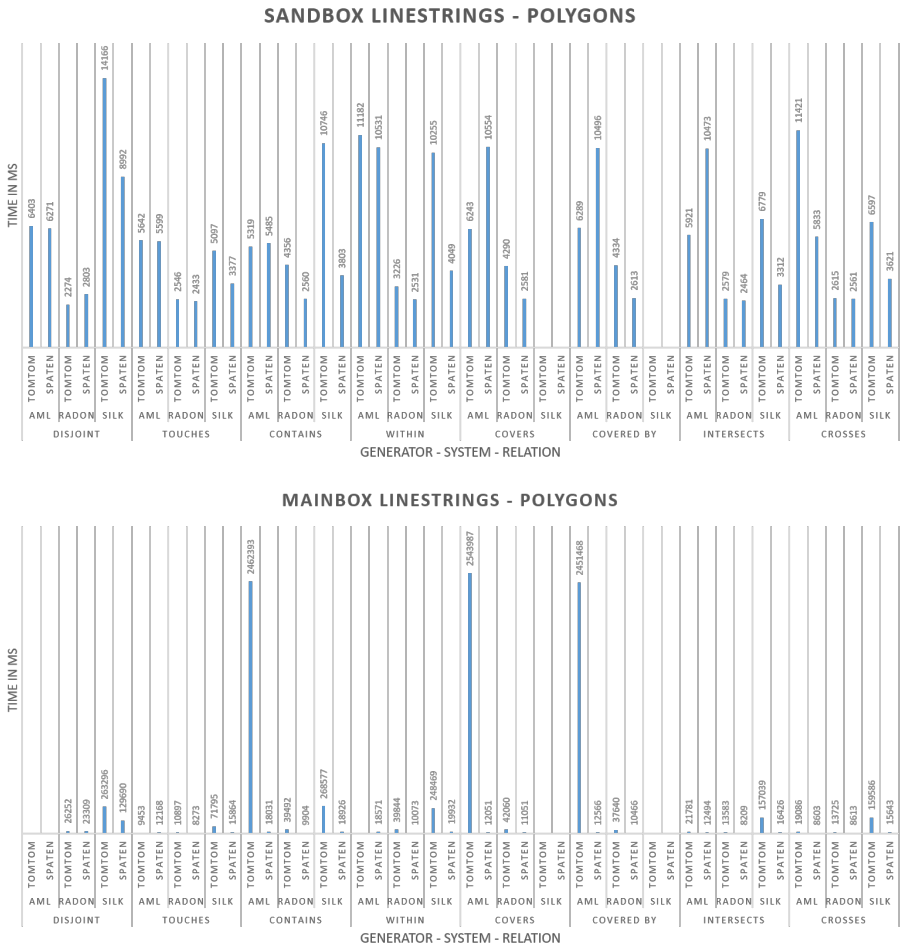

Fig. 3. Time performance for TomTom & Spaten SLP (top) and LLP (bottom) suites for AML (A), Silk (S) and RADON (R).

Table 15. SPIMBENCH track results.

System	Precision	Recall	F-measure	Time (ms)
Sandbox (100 instances)				
AML	0.835	0.896	0.865	6220
Lily	0.849	1.0	0.919	1960
LogMap	0.938	0.763	0.841	5887
Mainbox (5000 instances)				
AML	0.839	0.884	0.860	37190
Lily	0.855	1.0	0.922	3103
LogMap	0.893	0.709	0.791	23494

4.11 Knowledge Graph

We evaluated all SEALS participants in the OAEI (even those not registered for the track) on a very small matching example[20]. This revealed that not all systems were able to cope with the task, and in the end only the following systems were evaluated: AML, POMap++, Hontology, DOME, LogMap (in its KG version), LogMapBio, LogMapLt.

Of these systems, the following were able output results for all nine test cases: POMAP++, Holontology, DOME, LogMapBio and the baseline. AML ran out of time (12 hours) on some tracks, LogMap needed more than the given 32 GB RAM for the bigger knowledge graphs, and LogMapLt created alignment files bigger than 1GB (up to 50 GB in some runs).

Table 16 shows the aggregated results for each system, including the number of tasks in which it was able to generate a non-empty alignment (#tasks) and the average number of generated correspondences in those tasks (size). In addition to the global average precision, F-measure, and recall results, in which tasks where systems produced empty alignments were counted, we also computed F-measure and recall ignoring empty alignments (note that precision is the same) which are shown between parentheses in the table, where applicable.

All systems were able to generate class correspondences, but only the three tasks from the Games topic have enough classes to be meaningfully matched. The baseline has an F-Measure of 0.79 which is surpassed by AML, Holontology, LogMap and LogMapBio (when considering only completed tracks).

DOME was the only system able to produce property correspondences (in addition to the baseline). The remaining systems do not return any property correspondences, probably because all properties are typed as `rdf:Property` and not subdivided into `owl:DatatypeProperty` and `owl:ObjectProperty`. However, this cannot be done easily in a preprocessing step because the usage of the properties is not strict, i.e., some properties are used both with literals and resources as their object. Given that a system that matches only OWL properties of the same type would not be able to handle such cases as this, an improvement of these matching systems would be to include also the ability of correspondence `rdf:Property` in case no more types are defined.

[20] `http://oaei.ontologymatching.org/2018/results/knowledgegraph/small_test.zip`

Table 16. Knowledge Graph track results, divided into class, property, instance, and overall correspondences.

System	Time (s)	# tasks	Size	Prec.	F-m.	Rec.
Class performance						
AML	88448	5	11.6	0.85	0.64 (0.87)	0.51 (0.88)
POMAP++	438	9	15.1	0.79	0.74	0.69
Holontology	318	9	16.8	0.80	0.83	0.87
DOME	13747	9	16.0	0.73	0.73	0.73
LogMap	14083	7	21.7	0.66	0.77 (0.80)	0.91 (1.00)
LogMapBio	2340	9	22.1	0.68	0.81	1.00
LogMapLt	500	6	22.0	0.61	0.72 (0.76)	0.87 (1.00)
Baseline	412	9	18.9	0.75	0.79	0.84
Property performance						
AML	88448	5	0.0	0.00	0.00	0.00
POMAP++	438	9	0.0	0.00	0.00	0.00
Holontology	318	9	0.0	0.00	0.00	0.00
DOME	13747	9	207.3	0.86	0.84	0.81
LogMap	14083	7	0.0	0.00	0.00	0.00
LogMapBio	2340	9	0.0	0.00	0.00	0.00
LogMapLt	500	6	0.0	0.00	0.00	0.00
Baseline	412	9	213.8	0.86	0.84	0.82
Instance performance						
AML	88448	5	82380.9	0.16	0.23 (0.26)	0.38 (0.63)
POMAP++	438	9	0.0	0.00	0.00	0.00
Holontology	318	9	0.0	0.00	0.00	0.00
DOME	13747	9	15688.7	0.61	0.61	0.61
LogMap	14083	7	97081.4	0.08	0.14 (0.15)	0.81 (0.93)
LogMapBio	2340	9	0.0	0.00	0.00	0.00
LogMapLt	500	6	82388.3	0.39	0.52 (0.56)	0.76 (0.96)
Baseline	412	9	17743.3	0.59	0.69	0.82
Overall performance						
AML	88448	5	102471.1	0.19	0.23 (0.28)	0.31 (0.52)
POMAP++	438	9	16.9	0.79	0.14	0.08
Holontology	318	9	18.8	0.80	0.17	0.10
DOME	13747	9	15912.0	0.68	0.68	0.67
LogMap	14083	7	97104.8	0.09	0.16 (0.16)	0.64 (0.74)
LogMapBio	2340	9	24.1	0.68	0.19	0.11
LogMapLt	500	6	88893.1	0.42	0.49 (0.54)	0.60 (0.77)
Baseline	412	9	17976.0	0.65	0.73	0.82

With respect to instance correspondences, AML, DOME, LogMap, LogMapLt were able to produce them (as was the baseline) whereas POMAP++, Holontology and LogMapBio were not, since they are not designed for instance matching. The baseline was unsurpassed by any system in this category in either F-measure or recall. One reason for this is that the baseline had the highest F-measure among systems able to match both classes and instances, and had a higher F-measure than DOME at matching

properties, given that the alignment of instances is conditioned by the correct alignment of classes and properties. Furthermore, many of the matching systems return n:m correspondences and thus a lot of false positive correspondences, resulting in low precision.

We analyzed the errors for a specific task, namely `darkscape-oldschoolrunescape`. For this task, the baseline could not find the following correspondences: `Lumbridge_and_Draynor_Tasks = Lumbridge_&_Draynor_Diary` and `Cupric_sulphate = Cupric_sulfate`. The matcher AML does not find `Translated_notes = Translated_notes`, even if the label (wiki page name) is exactly the same. False positive correspondences for the LogMap matcher are `Ancient_Magicks = Carrallangar_Teleport` and `Ancient_Magicks = Kharyrll_Teleport`. For AML one example is `Customs_Officer = Gang_boss`.

Regarding runtime, AML was the slowest system, followed by DOME and LogMap. POMAP++ and Holontology were quite fast, but only return class correspondences.

4.12 Interactive matching

This year, the same four systems as last year participated in the Interactive matching track: ALIN, AML, LogMap, and XMap. Their results are shown in Table 17 and Figure 4 for both Anatomy and Conference datasets.

The table includes the following information (column names within parentheses):

- 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. 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 correspondences, that could be analysed simultaneously by a user.
- Distinct correspondences (Dist. Mapps) counts the total number of correspondences 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).

The figure shows the time intervals between the questions to the user/oracle for the different systems and error rates. Different runs are depicted with different colors.

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	0.998	0.611	0.758	0.0	–	–	–	–	–	–	–
	0.0	0.994	0.826	0.902	0.543	0.994	0.826	0.902	602	1448	1.0	1.0
	0.1	0.914	0.802	0.854	0.482	0.994	0.833	0.906	578	1373	0.731	0.965
	0.2	0.848	0.784	0.815	0.436	0.994	0.839	0.91	564	1343	0.561	0.931
	0.3	0.784	0.757	0.77	0.369	0.995	0.843	0.912	552	1307	0.419	0.875
AML	NI	0.95	0.936	0.943	0.832	–	–	–	–	–	–	–
	0.0	0.964	0.948	0.956	0.862	0.964	0.948	0.956	240	240	1.0	1.0
	0.1	0.952	0.946	0.948	0.857	0.965	0.95	0.957	268	268	0.719	0.97
	0.2	0.938	0.941	0.939	0.849	0.965	0.95	0.957	272	272	0.52	0.935
	0.3	0.92	0.938	0.929	0.843	0.966	0.951	0.958	299	299	0.379	0.905
LogMap	NI	0.918	0.846	0.88	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	0.961	0.832	0.892	0.568	0.964	0.801	0.875	388	1164	0.742	0.966
	0.2	0.945	0.823	0.88	0.552	0.944	0.761	0.842	388	1164	0.567	0.927
	0.3	0.932	0.819	0.872	0.543	0.922	0.725	0.812	388	1164	0.434	0.878
XMap	NI	0.929	0.865	0.896	0.647	–	–	–	–	–	–	–
	0.0	0.929	0.867	0.897	0.653	0.929	0.867	0.897	35	35	1.0	1.0
	0.1	0.929	0.867	0.897	0.653	0.929	0.866	0.896	35	35	0.601	0.978
	0.2	0.929	0.867	0.897	0.653	0.929	0.865	0.896	35	35	0.4	0.965
	0.3	0.929	0.867	0.897	0.653	0.929	0.863	0.895	35	35	0.298	0.946
Conference Dataset												
ALIN	NI	0.88	0.456	0.601	–	–	–	–	–	–	–	–
	0.0	0.921	0.721	0.809	–	0.921	0.721	0.809	276	698	1.0	1.0
	0.1	0.725	0.686	0.705	–	0.934	0.753	0.834	264	674	0.538	0.987
	0.2	0.601	0.648	0.623	–	0.942	0.773	0.849	260	657	0.341	0.967
	0.3	0.495	0.624	0.552	–	0.951	0.796	0.866	259	645	0.226	0.95
AML	NI	0.841	0.659	0.739		–	–	–	–	–	–	–
	0.0	0.912	0.711	0.799	–	0.912	0.711	0.799	270	270	1.0	1.0
	0.1	0.838	0.698	0.762	–	0.923	0.733	0.817	277	277	0.691	0.971
	0.2	0.769	0.676	0.719	–	0.928	0.747	0.827	271	271	0.533	0.922
	0.3	0.715	0.663	0.688	–	0.931	0.758	0.836	270	270	0.459	0.885
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	0.85	0.596	0.7	–	0.858	0.576	0.69	82	246	0.71	0.978
	0.2	0.82	0.588	0.685	–	0.831	0.547	0.66	82	246	0.507	0.941
	0.3	0.793	0.583	0.672	–	0.808	0.518	0.631	82	246	0.366	0.907
XMap	NI	0.716	0.62	0.665	–	–	–	–	–	–	–	–
	0.0	0.719	0.62	0.666	–	0.719	0.62	0.666	16	16	0.0	1.0
	0.1	0.719	0.62	0.666	–	0.719	0.617	0.665	16	16	0.0	1.0
	0.2	0.718	0.62	0.666	–	0.72	0.613	0.662	16	16	0.2	1.0
	0.3	0.718	0.62	0.666	–	0.721	0.613	0.662	16	16	0.1	1.0

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

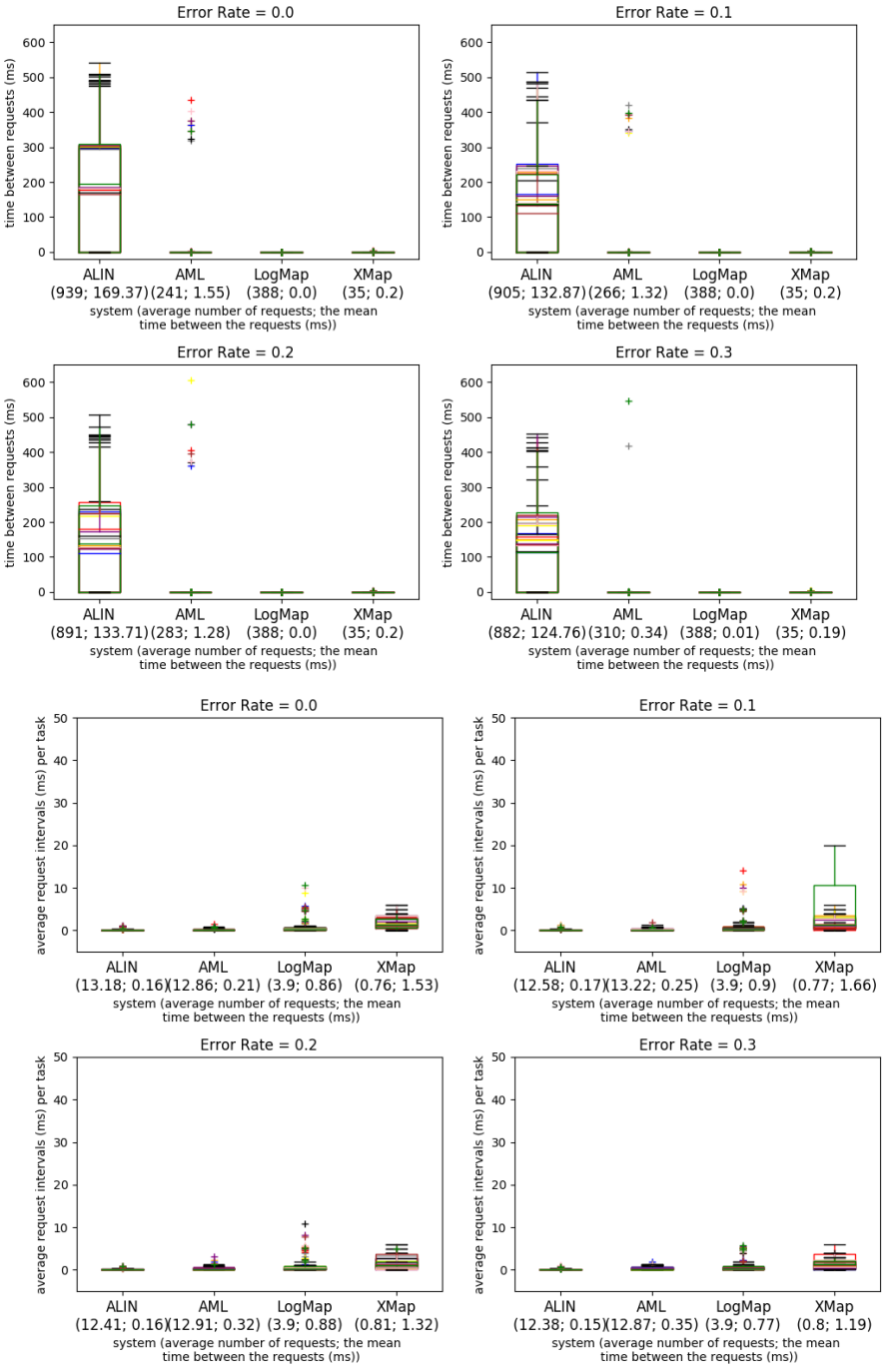

Fig. 4. Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

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 correspondences, ALIN can also add new candidate correspondences to its initial set. LogMap and AML both request feedback on only selected correspondences candidates (based on their similarity patterns or their involvement in unsatisfiabilities) and AML presents one correspondence at a time to the user. XMap also presents one correspondence at a time and asks mainly about incorrect correspondences. ALIN and LogMap can both ask the oracle to analyze several conflicting correspondences simultaneously.

The performance of the systems usually improves when interacting with a perfect oracle in comparison with no interaction. The one exception is XMap, because it is barely interactive in the datasets. In general, XMap performs very few requests to the oracle compared to the other systems. 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, because its high number of oracle requests and its non-interactive performance was the lowest of the interactive systems, and thus the easiest to improve.

Although system 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 the oracle's 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 datasets.

Another aspect that was assessed, was the response time of systems, i.e., the time between requests. Two models for system *response times* are frequently used in the literature [9]: Shneiderman and Seow take different approaches to categorize the response times taking a task-centered view and a user-centered view respectively. According to task complexity, Shneiderman defines response time in four categories: typing, mouse movement (50-150 ms), simple frequent tasks (1 s), common tasks (2-4 s) and complex tasks (8-12 s). While Seow's definition of response time is based on 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 datasets 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 datasets. 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).

4.13 Complex Matching

The only systems able to generate any kind of complex correspondence in any of the complex matching test cases were AMLC (in the Conference test suite) and CANARD

(in the Taxon test case). No systems were capable of generating complex correspondences over either the Hydrography or the GeoLink test cases.

On the Conference test suite, only complex correspondences were being evaluated, since simple correspondences are already evaluated under the Conference track. In the case of the Hydrography and GeoLink test cases, all SEALS OAEI participants were evaluated in subtask 1 of both test cases, wherein they had to simply identify related entities. On the Taxon test case, all 14 systems which registered to the complex, conference and/or anatomy track were evaluated, but only 7 could output at least one alignment.

The results of the systems on the four test cases are summarized in Table 18.

Table 18. Results of the Complex Track. The precision, recall and F-measure are the average measures. QWR is the proportion of queries well rewritten.

Matcher	Conference			Hydrography (subtask 1)			GeoLink (subtask 1)			Taxon	
	Prec.	F-meas.	Rec.	Prec.	F-meas.	Rec.	Prec.	F-meas.	Rec.	Prec.	QWR
ABC	-	-	-	0.43	**0.18**	**0.12**	-	-	-	-	-
ALOD2Vec	-	-	-	0.5	0.09	0.05	0.78	**0.19**	**0.11**	-	-
AMLC	**0.54**	**0.42**	**0.34**	-	-	-	-	-	-	-	-
AML	-	-	-	-	-	-	-	-	-	0.00	0.00
CANARD	-	-	-	-	-	-	-	-	-	0.20	**0.13**
DOME	-	-	-	0.35	0.09	0.06	0.44	0.17	**0.11**	-	-
FMapX	-	-	-	**0.46**	0.11	0.07	-	-	-	-	-
Holontology	-	-	-	-	-	-	-	-	-	0.22	0.00
KEPLER	-	-	-	0.5	0.09	0.05	-	-	-	-	-
LogMap	-	-	-	0.44	0.08	0.05	0.85	0.18	0.1	**0.54**	0.07
LogMapBio	-	-	-	-	-	-	-	-	-	0.28	0.00
LogMapKG	-	-	-	-	-	-	0.85	0.18	0.1	-	-
LogMapLt	-	-	-	-	-	-	0.73	**0.19**	**0.11**	0.16	0.10
POMAP++	-	-	-	0.42	0.06	0.04	**0.9**	0.17	0.09	0.14	0.00
XMap	-	-	-	0.21	0.09	0.06	0.39	0.15	0.09	-	-

With respect to subtask 1 of the Hydrography and GeoLink test cases, the results show that a simple baseline approach that identifies target entity names within source entity comments performs better than most existing matchers. This is unsurprising, as matching systems are configured to find equivalent concepts rather than related ones. The takeaway from this year is that there is a lot of room for new approaches on this task.

In the Taxon test cases, only the output of LogMap, LogMapLt and CANARD could be used to rewrite source queries.

A more detailed discussion of the results of each task can be found in the OAEI page for this track. For a first edition of complex matching in an OAEI campaign, and given the inherent difficulty of the task, the results and participation are promising albeit still modest.

5 Conclusions & Lessons Learned

The OAEI 2018 counted this year several new tracks, some of which open new perspectives in the field, in particular with respect to the generation of more expressive alignments. We witnessed a slight decrease in the number of participants in comparison with previous years, but with a healthy mix of new and returning systems. However, like last year, the distribution of participants by tracks was uneven.

The **schema matching tracks** saw abundant participation, but, as has been the trend of the recent years, little substantial progress in terms of quality of the results or run time of top matching systems, judging from the long-standing tracks. On the one hand, this may be a sign of a performance plateau being reached by existing strategies and algorithms, which would suggest that new technology is needed to obtain significant improvements. On the other hand, it is also true that established matching systems tend to focus more on new tracks and datasets than on improving their performance in long-standing tracks, whereas new systems typically struggle to compete with established ones.

The number of matching systems capable of handling very large ontologies has increased slightly over the last years, but is still relatively modest, judging from the *Large Biomedical Ontologies* track. We will aim at facilitating participation in future editions of this track by providing techniques to divide the matching tasks in manageable subtasks (e.g., [27]).

There has also been progress, but likewise room for improvement, on the ability of matching systems to match properties, judging from the *Conference* track. To assist system developers in tackling this aspect, we plan to provide a more detailed evaluation in the future, including an analysis of the false positives per matching system.

Less encouraging is the low number of systems concerned with the logical coherence of the alignments they produce, an aspect which is critical for several semantic web applications. Perhaps a more direct approach is needed to promote this topic, such as providing a more in-depth analysis of the causes of incoherence in the evaluation or even organizing a future track focusing on logical coherence alone.

The consensus-based evaluation in the *Disease and Phenotype* track offers limited insights into performance, as several matching systems produce a number of unique correspondences which may or may not be correct. In the absence of a true reference alignment, future evaluation should seek to determine whether the unique correspondences contain indicators of correctness, such as semantic similarity, or appear to be noise.

The **instance matching tracks** and the new **instance and schema matching track** counted few participants, as has been the trend in recent years. Part of the reason for this is that several of these tracks ran on the HOBBIT platform, and the transition from SEALS to HOBBIT has not been as easy as we might desire. Thus, participation should increase next year as systems become more familiar with the HOBBIT platform and have more time to do the migration. Furthermore, from an infrastructure point of view, the HOBBIT SDK will make the developing and debugging phase easier, and the Maven-based framework will facilitate submission. However, another factor behind the reduced participation in the instance matching tracks lies with their specialization. New schema matching tracks such as *Biodiversity and Ecology* typically demand very

little from systems that are already able to tackle long-standing tracks such as *Anatomy*, whereas instance matching tracks such as *IIMB*, *Link Discovery* and last year's *Process Model Matching*, are so different from one another that each requires dedicated development time to tackle. Thus, in future OAEI editions we should consider publishing new instance matching (and other more specialized) datasets with more time in advance, to give system developers adequate time to tackle them. Equally critical will be to ensure stability by maintaining instance matching tracks and datasets over multiple OAEI editions, so that participants can build upon the development of previous years.

Automatic instance-matching benchmark generation algorithms have been gaining popularity, as evidenced by the fact that they are used in all three instance matching tracks of this OAEI edition. One aspect that has not been addressed in such algorithms is that, if the transformation is too extreme, the correspondence may be unrealistic and impossible to detect even by humans. As such, we argue that *human-in-the-loop* techniques can be exploited to do a preventive quality-checking of generated correspondences, and refine the set of correspondences included in the final reference alignment. We will explore such an approach in future editions of the *IIMB* track.

The **interactive matching track** also witnessed a small number of participants, which have been the same 4 systems over the last three campaigns. This is puzzling considering that this track is based on the *Anatomy* and *Conference* test cases, and those tracks had 14 participants. The process of programmatically querying the Oracle class used to simulate user interactions is simple enough that it should not be a deterrent for participation, but perhaps we should look at facilitating the process further in future OAEI editions by providing implementation examples.

Finally, the **complex matching track** opens new perspectives in the field of ontology matching, as this is a topic largely unexplored but of growing importance, since integrating linked datasets often encompasses making complex correspondences. Tackling complex matching automatically is extremely challenging, likely requiring profound adaptations from matching systems, so the fact that there were two participants able to generate complex correspondences in this track should be seen as a positive sign of progress to the state of the art in ontology matching. While this year the track involved different evaluation settings, we will work towards enabling the automatic evaluation of complex alignments in future editions.

Like in previous OAEI editions, most participants provided a description of their systems and their experience in the evaluation, in the form of OAEI system papers. These papers, like the present one, have not been peer reviewed. However, they are full contributions to this evaluation exercise, reflecting the effort and insight of matching systems developers, and providing details about those systems and the algorithms they implement.

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 [41]. 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 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 dataset.

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

We thank Andrea Turbati and the AGROVOC team for their very appreciated help with the preparation of the AGROVOC subset ontology. We are also grateful to Catherine Roussey and Nathalie Hernandez for their help on the Taxon alignment.

We also thank for their support the past members of the Ontology Alignment Evaluation Initiative steering committee: Jérôme Euzenat (INRIA, FR), 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).

Cássia Trojahn dos Santos has been partially supported by the French CIMI Labex project IBLiD (Integration of Big and Linked Data for On-Line Analytics).

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

Ernesto Jimenez-Ruiz has been partially supported by the BIGMED project (IKT 259055), the SIRIUS Centre for Scalable Data Access (Research Council of Norway, project no.: 237889), and the AIDA project (UK Government's Defence & Security Programme in support of the Alan Turing Institute.

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.

Irini Fundulaki and Tzanina Saveta were supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 688227 (Hobbit).

Jana Vataščinová and Ondřej Zamazal have been supported by the CSF grant no. 18-23964S.

Patrick Lambrix and Huanyu Li have been supported by the Swedish e-Science Research Centre and the Swedish National Graduate School in Computer Science (CUGS).

References

1. Manel Achichi, Michelle Cheatham, Zlatan Dragisic, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Ian Harrow, Valentina Ivanova, Ernesto Jiménez-

Ruiz, Kristian Kolthoff, Elena Kuss, Patrick Lambrix, Henrik Leopold, Huanyu Li, Christian Meilicke, Majid Mohammadi, Stefano Montanelli, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Heiner Stuckenschmidt, Élodie Thiéblin, Konstantin Todorov, Cássia Trojahn, and Ondrej Zamazal. Results of the ontology alignment evaluation initiative 2017. In *Proceedings of the 12th International Workshop on Ontology Matching, Vienna, Austria, October 21*, pages 61–113, 2017.

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 *Proceedings of the 11th International Ontology matching workshop, Kobe (JP), October 18th*, 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 *Proceedings of the 7th International Ontology matching workshop, Boston (MA, US)*, pages 73–115, 2012.
4. Benhamin Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Proc. K-Cap Workshop on Integrating Ontologies*, Banff (Canada), 2005.
5. Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.
6. 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 *Proceedings of the 3rd Ontology matching workshop, Karlsruhe (DE)*, pages 73–120, 2008.
7. Michelle Cheatham, Zlatan Dragisic, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, Patrick Lambrix, Stefano Montanelli, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Alessandro Solimando, Cássia Trojahn, and Ondřej Zamazal. Results of the ontology alignment evaluation initiative 2015. In *Proceedings of the 10th International Ontology matching workshop, Bethlehem (PA, US)*, pages 60–115, 2015.
8. 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, *Proceedings of the 8th International Ontology matching workshop, Sydney (NSW, AU)*, pages 61–100, 2013.
9. 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.
10. Thaleia Dimitra Doudali, Ioannis Konstantinou, and Nectarios Koziris Doudali. Spaten: a Spatio-Temporal and Textual Big Data Generator. In *IEEE Big Data*, pages 3416–3421, 2017.
11. 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 on-

tology alignment evaluation initiative 2014. In *Proceedings of the 9th International Ontology matching workshop, Riva del Garda (IT)*, pages 61–104, 2014.
12. Zlatan Dragisic, Valentina Ivanova, Patrick Lambrix, Daniel Faria, Ernesto Jiménez-Ruiz, and Catia Pesquita. User validation in ontology alignment. In *Proceedings of the 15th International Semantic Web Conference, Kobe, Japan, October 17-21*, pages 200–217, 2016.
13. Zlatan Dragisic, Valentina Ivanova, Huanyu Li, and Patrick Lambrix. Experiences from the anatomy track in the ontology alignment evaluation initiative. *Journal of Biomedical Semantics*, 8:56:1–56:28, 2017.
14. 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 *Proceedings of the 4th International Ontology matching workshop, Chantilly (VA, US)*, pages 73–126, 2009.
15. 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 *Proceedings of the 5th International Ontology matching workshop, Shanghai (CN)*, pages 85–117, 2010.
16. 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 *Proceedings of the 6th International Ontology matching workshop, Bonn (DE)*, pages 85–110, 2011.
17. 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 *Proceedings 2nd International Ontology matching workshop, Busan (KR)*, pages 96–132, 2007.
18. 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.
19. 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 *Proceedings of the 1st International Ontology matching workshop, Athens (GA, US)*, pages 73–95, 2006.
20. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition, 2013.
21. Daniel Faria, Ernesto Jiménez-Ruiz, Catia Pesquita, Emanuel Santos, and Francisco M. Couto. Towards Annotating Potential Incoherences in BioPortal Mappings. In *Proceedings of the 13th International Semantic Web Conference*, volume 8797, pages 17–32. Springer, 2014.
22. Alfio Ferrara, Stefano Montanelli, Jan Noessner, and Heiner Stuckenschmidt. Benchmarking matching applications on the semantic web. In *Proceedings of the 8th Extended Semantic Web Conference*, Heraklion, Greece, 2011.
23. 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*, 2017.
24. Sven Hertling and Heiko Paulheim. Dbkwik: A consolidated knowledge graph from thousands of wikis. In *Proceedings of the International Conference on Big Knowledge*, 2018.

25. Alexandra Hofmann, Samresh Perchani, Jan Portisch, Sven Hertling, and Heiko Paulheim. Dbkwik: Towards knowledge graph creation from thousands of wikis. In *Proceedings of the International Semantic Web Conference (Posters and Demos), Vienna, Austria*, pages 21–25, 2017.
26. Valentina Ivanova, Patrick Lambrix, and Johan Åberg. Requirements for and evaluation of user support for large-scale ontology alignment. In *Proceedings of the European Semantic Web Conference*, pages 3–20, 2015. doi: 10.1007/978-3-319-18818-8_1.
27. Ernesto Jiménez-Ruiz, Asan Agibetov, Matthias Samwald, and Valerie Cross. Breaking-down the Ontology Alignment Task with a Lexical Index and Neural Embeddings. *CoRR*, abs/1805.12402, 2018.
28. Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. LogMap: Logic-based and scalable ontology matching. In *Proceedings of the 10th International Semantic Web Conference, Bonn (DE)*, pages 273–288, 2011.
29. 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.
30. Ernesto Jiménez-Ruiz, Christian Meilicke, Bernardo Cuenca Grau, and Ian Horrocks. Evaluating mapping repair systems with large biomedical ontologies. In *Proceedings of the 26th Description Logics Workshop*, 2013.
31. Ernesto Jiménez-Ruiz, Tzanina Saveta, Ondvrej Zamazal, Sven Hertling, Michael Röder, Irini Fundulaki, Axel-Cyrille Ngonga Ngomo, Mohamed Ahmed Sherif, Amina Annane, Zohra Bellahsene, Sadok Ben Yahia, Gayo Diallo, Daniel Faria, Marouen Kachroudi, Abderrahmane Khiat, Patrick Lambrix, Huanyu Li, Maximilian Mackeprang, Majid Mohammadi, Maciej Rybinski, Booma Sowkarthiga Balasubramani, and Cassia Trojahn. Introducing the HOBBIT platform into the Ontology Alignment Evaluation Campaign. In *Proceedings of the 13th International Workshop on Ontology Matching*, 2018.
32. Naouel Karam, Claudia Müller-Birn, Maren Gleisberg, David Fichtmüller, Robert Tolksdorf, and Anton Güntsch. A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data. *Datenbank-Spektrum*, 16(3):195–205, 2016.
33. Yevgeny Kazakov, Markus Krötzsch, and Frantisek Simancik. Concurrent classification of EL ontologies. In *Proceedings of the 10th International Semantic Web Conference, Bonn (DE)*, pages 305–320, 2011.
34. Christian Meilicke. *Alignment Incoherence in Ontology Matching*. PhD thesis, University Mannheim, 2011.
35. 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.
36. Boris Motik, Rob Shearer, and Ian Horrocks. Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009.
37. Heiko Paulheim, Sven Hertling, and Dominique Ritze. Towards evaluating interactive ontology matching tools. In *Proceedings of the 10th Extended Semantic Web Conference, Montpellier (FR)*, pages 31–45, 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 M Couto. Ontology alignment repair through modularization and confidence-based heuristics. *PLoS ONE*, 10(12):e0144807, 2015.

40. Tzanina Saveta, Evangelia Daskalaki, Giorgos Flouris, Irini Fundulaki, Melanie Herschel, and Axel-Cyrille Ngonga Ngomo. Pushing the limits of instance matching systems: A semantics-aware benchmark for linked data. In *Proceedings of the 24th International Conference on World Wide Web*, pages 105–106, New York, NY, USA, 2015. ACM.
41. 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.
42. Alessandro Solimando, Ernesto Jiménez-Ruiz, and Giovanna Guerrini. Detecting and correcting conservativity principle violations in ontology-to-ontology mappings. In *Proceedings of the International Semantic Web Conference*, pages 1–16. Springer, 2014.
43. Alessandro Solimando, Ernesto Jimenez-Ruiz, and Giovanna Guerrini. Minimizing conservativity violations in ontology alignments: Algorithms and evaluation. *Knowledge and Information Systems*, 2016.
44. Christian Strobl. *Encyclopedia of GIS*, chapter Dimensionally Extended Nine-Intersection Model (DE-9IM), pages 240–245. Springer, 2008.
45. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the Workshop on Evaluation of Ontology-based Tools (EON), Hiroshima (JP)*, 2004.
46. Elodie Thiéblin, Michelle Cheatham, Cassia Trojahn, Ondrej Zamazal, and Lu Zhou. The First Version of the OAEI Complex Alignment Benchmark. In *Proceedings of the International Semantic Web Conference (Posters and Demos)*, 2018.
47. Ondřej Zamazal and Vojtěch Svátek. The ten-year ontofarm and its fertilization within the onto-sphere. *Web Semantics: Science, Services and Agents on the World Wide Web*, 43:46–53, 2017.
48. Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. A complex alignment benchmark: Geolink dataset. In *Proceedings of the 17th International Semantic Web Conference, Monterey (CA, USA), October 8-12*, pages 273–288, 2018.

Jena, Dayton, Lisboa, Milano, Heraklion,
Mannheim, Oslo, Berlin, Linköping, Porto Alegre,
Trento, Toulouse, Prague, London
November 2018