Results of the Ontology Alignment Evaluation Initiative 2019*

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

[1] Friedrich Schiller University Jena, Germany
alsayed.algergawy@uni-jena.de
[2] BioData.pt, INESC-ID, Lisbon, Portugal
dfaria@inesc-id.pt
[3] Università degli studi di Milano, Italy
{alfio.ferrara,stefano.montanelli}@unimi.it
[4] Institute of Computer Science-FORTH, Heraklion, Greece
{jsaveta,fundul}@ics.forth.gr
[5] Pistoia Alliance Inc., USA
{ian.harrow,andrea.splendiani}@pistoiaalliance.org
[6] University of Mannheim, Germany
{sven,heiko}@informatik.uni-mannheim.de
[7] City, University of London, UK
ernesto.jimenez-ruiz@city.ac.uk
[8] Department of Informatics, University of Oslo, Norway
ernestoj@ifi.uio.no
[9] Fraunhofer FOKUS, Berlin, Germany
naouel.karam@fokus.fraunhofer.de
[10] Fraunhofer IAIS, Sankt Augustin, Bonn, Germany
abderrahmane.khiat@iais.fraunhofer.de
[11] Linköping University & Swedish e-Science Research Center, Linköping, Sweden
{patrick.lambrix,huanyu.li}@liu.se
[12] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
cpesquita@di.fc.ul.pt
[13] TasLab, Trentino Digitale SpA, Trento, Italy
pavel.shvaiko@tndigit.it
[14] IRIT & Université Toulouse II, Toulouse, France
{cassia.trojahn,elodie.thieblin}@irit.fr
[15] University of Economics, Prague, Czech Republic
{jana.vatascinova,ondrej.zamazal}@vse.cz
[16] Data Semantics (DaSe) Laboratory, Kansas State University, USA
luzhou@ksu.edu

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 2019 campaign offered 11 tracks with 29 test cases, and was attended by 20 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 [21, 23]. 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) [48]. Then, a unique OAEI campaign occurred in 2005 at the workshop on Integrating Ontologies held in conjunction with the International Conference on Knowledge Capture (K-Cap) [5]. From 2006 until the present, the OAEI campaigns were held at the Ontology Matching workshop, collocated with ISWC [4, 1–3, 7, 8, 10, 13, 17–20, 22], which this year took place in Auckland, New Zealand[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 2019 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 discusses the lessons learned.

[1] http://oaei.ontologymatching.org
[2] http://om2019.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 includes a SystemAdapter class, then being uploaded into the HOBBIT platform [34].

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 2019 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}, 2019. 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 30^{th}, 2019, 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 the 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 OAEI web pages by October 14^{th}, 2019.

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

3 Tracks and test cases

This year's OAEI campaign consisted of 11 tracks gathering 29 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 (2445)	=	[0 1]	open+blind	AR, CZ, CN, DE, EN, ES, FR, IT, NL, RU, PT	SEALS
Instance Matching						
Link Discovery	2 (9)	=	[0 1]	open	EN	HOBBIT
SPIMBENCH	2	=	[0 1]	open+blind	EN	HOBBIT
Instance and Schema Matching						
Knowledge Graph	5	=	[0 1]	open	EN	SEALS
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 [15].

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, correspondences 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 second edition of biodiversity track features two test cases based on highly overlapping ontologies that are particularly useful for biodiversity and ecology research: matching Environment Ontology (ENVO) to Semantic Web for Earth and Environment Technology Ontology (SWEET), and matching Flora Phenotype Ontology (FLOPO) to 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 [35, 37]. Table 2 summarizes the versions and the sizes of the ontologies used in OAEI 2019. Compared to the first edition, the number of concepts of the ENVO and FOLPO ontologies has increased, which required the creation of new reference alignments for both tasks.

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

Ontology	Version	Classes
ENVO	2019-03-18	8968
SWEET	2018-03-12	4543
FLOPO	2016-06-03	28965
PTO	2017-09-11	1504

[5] 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

To this end, we updated the reference alignments for the two test cases following the same procedure as in the first edition. In particular, alignment files were produced through a hybrid approach consisting of (1) an updated consensus alignment based on matching systems output, then (2) manually validating a subset of unique mappings produced by each system (and adding them to the consensus if considered correct), and finally (3) adding a set of manually generated correspondences. The matching systems used to generate the consensus alignments were those participating in this track last year [4], namely: AML, Lily, LogMap family, POMAP and XMAP.

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 [52].

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 correspondence 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 used 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 [25]. Table 3.4 summarizes the versions of the ontologies used in OAEI 2019.

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 [41]. The evaluation was carried out in a Ubuntu 18 Laptop with an Intel Core i5-6300HQ CPU @ 2.30GHz x 4 and allocating 15 Gb of 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 [6] as detailed in [32], 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 [39], LogMap [31], or AML [43]. 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 i5-6300HQ CPU @ 2.30GHz x 4 and allocating 15 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 [41], 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 OWL2 EL reasoner ELK [36].

3.6 Multifarm

The multifarm track [40] 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. The systems have been executed on a Ubuntu Linux machine configured with 8GB of RAM running under a Intel Core CPU 2.00GHz x4 processors, using the SEALS client.

3.7 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 [12].

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 [44], 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. 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 a ground truth was built that contains 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 [47]. 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. The focus for the design was (a) on the correct implementation of all the topological relations of the DE-9IM topological model and (b) on producing 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.

3.8 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,

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

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

blog post or programme). The datasets were generated and transformed using SPIMBENCH [44] 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 counterpart in the target dataset. The SPIMBENCH task uses two sets of 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).
- 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 cases, 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.

3.9 Knowledge Graph

The Knowledge Graph track was run for the second year. The task of the track is to match pairs of knowledge graphs, whose schema and instances have to be matched simultaneously. The individual knowledge graphs are created by running the DBpedia extraction framework on eight different Wikis from the Fandom Wiki hosting platform[13] in the course of the DBkWik project [27, 26]. They cover different topics (movies, games, comics and books) and three Knowledge Graph clusters shares the same domain e.g. star trek, as shown in Table 5.

The evaluation is based on reference correspondences at both schema and instance levels. While the schema level correspondences were created by experts, the instance correspondences were extracted from the wiki page itself. Due to the fact that not all inter wiki links on a page represent the same concept a few restrictions were made: 1) Only links in sections with a header containing "link" are used 2) all links are removed where the source page links to more than one concept in another wiki (ensures the alignments are functional) 3) multiple links which point to the same concept are also removed (ensures injectivity). 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_212, using the SEALS client (version 7.0.5). We used the $-\circ$ option in SEALS to provide the

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

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

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

Source	Hub	Topic	#Instances	#Properties	#Classes
Star Wars Wiki	Movies	Entertainment	145,033	700	269
The Old Republic Wiki	Games	Gaming	4,180	368	101
Star Wars Galaxies Wiki	Games	Gaming	9,634	148	67
Marvel Database	Comics	Comics	210,996	139	186
Marvel Cinematic Universe	Movies	Entertainment	17,187	147	55
Memory Alpha	TV	Entertainment	45,828	325	181
Star Trek Expanded Universe	TV	Entertainment	13,426	202	283
Memory Beta	Books	Entertainment	51,323	423	240

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 "-x" option of SEALS because the evaluation routine needed to be changed for two reasons: first, to differentiate between results for class, property, and instance correspondences, and second, to deal with the partial nature of the gold standard.

The alignments were evaluated based on precision, recall, and f-measure for classes, properties, and instances (each in isolation). The partial gold standard contained 1:1 correspondences and 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[14].

As a baseline, we employed two simple string matching approaches. The source code for these matchers is publicly available[15].

3.10 Interactive Matching

The interactive matching track aims to assess the performance of semi-automated matching systems by simulating user interaction [42, 14, 38]. 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 [29, 14].

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

[14] `http://oaei.ontologymatching.org/2019/results/knowledgegraph/matching-eval-trackspecific.zip`

[15] `http://oaei.ontologymatching.org/2019/results/knowledgegraph/kgBaselineMatchers.zip`

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).

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.

3.11 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 [51].

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 **populated complex conference** is a populated version of the Conference dataset. 5 ontologies have been populated with more or less common instances resulting in 6 datasets (6 versions on the seals repository: v0, v20, v40, v60, v80 and v100). The alignments were evaluated based on Competency Questions for Alignment, i.e., basic queries that the alignment should be able to cover [49]. The queries are automatically rewritten using 2 systems: that from [50] which covers (1:n) correspondences with EDOAL expressions; and a system which compares the answers (sets of instances or sets of pairs of instances) of the source query and the source member of the correspondences and which outputs the target member if both sets are identical. The best rewritten query scores are kept. A precision score is given by comparing the instances described by the source and target members of the correspondences.

The **Hydrography** dataset consists of matching four different source ontologies (hydro3, hydrOntology-translated, hydrOntology-native, and cree) to a single target ontology (SWO) [9]. 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 three subtasks were evaluated based on relaxed precision and recall [16].

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 the instance data is also available and populated into the benchmark. 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 [54, 55]. Evaluation was done in the same way as with the Hydrography dataset. The evaluation platform was a MacBook Pro with a 2.5 GHz Intel Core i7 processor and 16 GB of 1600 MHz DDR3 RAM running mac OS Yosemite version 10.10.5.

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, which is slightly over 20. This year we count with 20 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).

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.

Table 6. Participants and the status of their submissions.

System	AGM	ALIN	AML	AMLC	AROA	CANARD	DOME	EVOCROS	FCAMap-KG	FTRLIM	Lily	LogMap	LogMap-Bio	LogMapLt	OntMat1	POMAP++	RADON	SANOM	Silk	WktMtchr	Total=20
Confidence	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
anatomy	●	●	●	○	○	○	●	○	●	○	●	●	●	●	○	●	○	●	○	●	12
conference	○	●	●	○	○	○	●	○	○	○	●	●	○	●	●	○	○	●	○	●	9
multifarm	○	○	●	○	○	○	○	○	○	○	◐	●	○	○	○	○	○	○	○	●	4
complex	○	○	○	◐	◐	◐	○	○	○	○	○	○	○	○	○	○	○	○	○	○	3
interactive	○	●	●	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	3
largebio	●	○	●	○	○	○	●	○	●	○	○	●	●	●	○	●	○	●	○	●	10
phenotype	○	○	●	○	○	○	●	○	●	○	○	●	●	●	○	●	○	○	○	●	8
biodiv	○	○	●	○	○	○	●	○	●	○	○	●	●	●	○	●	○	○	○	○	7
spimbench	○	○	●	○	○	○	○	○	○	●	●	●	○	○	○	○	●	○	●	○	6
link discovery	○	○	●	○	○	○	○	○	○	●	●	○	○	○	○	○	●	○	●	○	6
knowledge graph	●	○	●	○	○	○	●	○	●	○	○	●	●	●	○	●	○	○	○	●	9
total	3	3	10	1	1	1	6	0	5	2	5	10	5	6	1	5	2	3	2	5	77

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.

4.2 Anatomy

The results for the Anatomy track are shown in Table 7. Of the 12 systems participating in the Anatomy track, 10 achieved an F-measure higher than the StringEquiv baseline. Two systems were first time participants (Wiktionary and AGM). Long-term participating systems showed few changes in comparison with previous years with respect to alignment quality (precision, recall, F-measure, and recall+), size and run time. The exceptions were LogMapBio which increased in both recall+ (from 0.756 to 0.801) and alignment size (by 57 correspondences) since last year, and ALIN that increased in F-measure (from 0.758 to 0.813) and recall+ (from 0.0 to 0.365), as well as had a substantial increase of 158 correspondences since last year.

In terms of run time, 5 out of 12 systems computed an alignment in less than 100 seconds, a ratio which is similar to 2018 (6 out of 14). 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 3 other systems obtained an F-measure above 0.88 (LogMapBio, POMap++, 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. Four systems produced coherent alignments.

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	76	1493	0.95	0.943	0.936	0.832	√
LogMapBio	1718	1607	0.872	0.898	0.925	0.801	√
POMAP++	345	1446	0.919	0.897	0.877	0.695	-
LogMap	28	1397	0.918	0.88	0.846	0.593	√
SANOM	516	-	0.888	0.865	0.844	0.632	-
Lily	281	1381	0.873	0.833	0.796	0.52	-
Wiktionary	104	1144	0.968	0.832	0.73	0.288	-
LogMapLite	19	1147	0.962	0.828	0.728	0.288	-
ALIN	5115	1086	0.974	0.813	0.698	0.365	√
FCAMap-KG	25	960	0.996	0.772	0.631	0.042	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
DOME	23	936	0.996	0.76	0.615	0.007	-
AGM	628	1942	0.152	0.171	0.195	0.154	-

4.3 Biodiversity and Ecology

Five of the systems participating this year had participated in this track in OAEI 2018: AML, LogMap family systems (LogMap, LogMapBio and LogMapLT) and POMAP. Three were new participants: DOME, FCAMapKG and LogMapKG. The newcomers DOME, FCAMapKG did not register explicitly to this track but could cope with at least one task so we did include their results.

We observed a slight increase in the number of systems (8 systems) that succeeded to generate alignments for the FLOPO-PTO task in comparison to previous year (7 systems). However, we witnessed a slight decrease in the number of systems (6 systems) that succeeded to generate alignments for the test ENVO-SWEET in comparison to previous year (7 systems). Lily did not manage to generate mappings for both tasks and LogMapBio did not manage to generated mappings for the ENVO-SWEET task.

As in the previous edition, we used precision, recall and F-measure to evaluate the performance of the participating systems. This year we included the execution times. The results for the Biodiversity and Ecology track are shown in Table 8.

Overall, the results of the participating systems have decreased in terms of F-measure for both tasks compared to last year. In terms of run time, most of the systems (except POMAP) computed an alignment in less than 100 seconds.

For the FLOPO-PTO task, AML and LogMapKG achieved the highest F-measure (0.78), with a slight difference in favor of AML. However, AML showed a remarkable decrease in terms of precision (from 0.88 to 0.76) and F-measure (from 0.86 to 0.78) compared to last year. LogMap also showed a slight decrease in terms of F-measure (from 0.80 to 0.78). The DOME system (newcomer) achieved the highest precision (0.99) with quite a good F-measure (0.739).

Regarding the ENVO-SWEET task, AML ranked first in terms of F-measure (0.80), followed by POMAP (0.69), FCAMapKG (0.63) and LogMapKG (0.63). As last year AML showed a very high recall and significant larger alignment than the other top

Table 8. Results for the Biodiversity & Ecology track.

System	Time (s)	Size	Precision	Recall	F-measure
FLOPO-PTO task					
AML	42	511	0.766	0.811	0.788
DOME	8.22	141	0.993	0.588	0.739
FCAMapKG	7.2	171	0.836	0.601	0.699
LogMap	14.4	235	0.791	0.782	0.768
LogMapBio	480.6	239	0.778	0.782	0.780
LogMapKG	13.2	235	0.791	0.782	0.786
LogMapLite	6.18	151	0.947	0.601	0.735
POMap	311	261	0.651	0.714	0.681
ENVO-SWEET task					
AML	3	925	0.733	0.899	0.808
FCAMapKG	7.8	422	0.803	0.518	0.630
LogMap	26.9	443	0.772	0.523	0.624
LogMapKG	7.98	422	0.803	0.518	0.630
LogMapLite	13.8	617	0.648	0.612	0.629
POMap	223	673	0.684	0.703	0.693

systems, but a comparably lower precision and a slight decrease in terms of F-measure (from 0.84 to 0.80). POMAP ranked second this year with a remarkable decrease in terms of precision (from 0.83 to 0.68) and F-measure (from 0.78 to 0.69). FCAMapKG and LogMapKG showed the highest results in terms of precision (0.80).

AML generated a significantly large number of mappings (much bigger than the size of the reference alignments for both tasks), those alignments were mostly subsumption mappings. In order to evaluate the precision in a more significant manner, we had to calculate an approximation by assessing manually a subset of mappings not present in the reference alignment (around a 100 for each task).

Overall, in this second evaluation, the results obtained from participating systems remained similar with a slight decrease in terms of F-measure compared to last year. It is worth noting 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 two baselines we can group tools according to matcher's position: four matching systems outperformed both baselines (SANOM, AML, LogMap and Wiktionary); two performed the same as the edna baseline (DOME and LogMapLt); one performed slightly worse than this baseline (ALIN); and two (Lily and ONTMAT1) performed worse than both baselines. Three matchers (ONTMAT1, ALIN and Lily) do

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
Wiktionary	0.65	0.62	0.58	0.54	0.52	7	133	27
DOME	0.73	0.65	0.56	0.5	0.46	3	105	10
edna	0.74	0.66	0.56	0.49	0.45			
LogMapLt	0.68	0.62	0.56	0.5	0.47	3	97	18
ALIN	0.81	0.68	0.55	0.46	0.42	0	2	0
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
ONTMAT1	0.77	0.64	0.52	0.43	0.39	1	71	37

not match properties at all. Naturally, this 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 [45, 46], only three tools (ALIN, AML and LogMap) have no consistency principle violation (the same tools as last year). This year all tools have some conservativity principle violations as the 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.

This year we additionally analyzed the False Positives, i.e. correspondences discovered by the tools which were evaluated as incorrect. The list of the False Positives is available on the conference track's web page. We looked at the reasons why a correspondence was incorrect or why it was discovered from a general point of view, and defined 3 reasons why alignments are incorrect and 5 reasons why they could have been chosen. Looking at the results, it can be said that when the reason a correspondence was discovered was the same name, all or at least most tools generated the correspondence. False Positives not discovered based on the same name or synonyms were produced by Lily, ONTMAT1 and SANOM. SANOM was the only tool which produced these correspondences based on similar strings. In three cases, a class was matched with a property by DOME (1x), LogMapLt (1x) and Wiktionary (3x).

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

Out of the 9 alignment systems, five (ALIN, DOME, LogMapLt, ONTMAT1, SANOM) use 1.0 as the confidence value for all matches they identify. The remaining

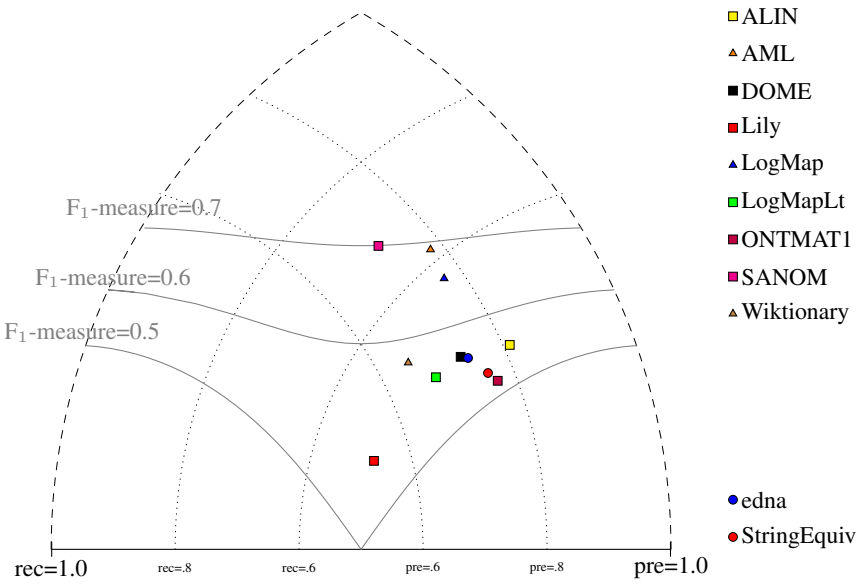

Fig. 1. Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F_1-measure are depicted by areas bordered by corresponding lines F_1-measure=0.[5|6|7].

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.
ALIN	0.87	0.58	0.44	0.87	0.68	0.56	0.87	0.69	0.57
AML	0.84	0.74	0.66	0.79	0.78	0.77	0.80	0.77	0.74
DOME	0.78	0.59	0.48	0.78	0.68	0.60	0.78	0.65	0.56
Lily	0.59	0.56	0.53	0.67	0.02	0.01	0.59	0.32	0.22
LogMap	0.82	0.69	0.59	0.81	0.70	0.62	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
ONTMAT1	0.82	0.55	0.41	0.82	0.64	0.52	0.82	0.64	0.53
SANOM	0.79	0.74	0.69	0.66	0.74	0.83	0.65	0.72	0.81
Wiktionary	0.69	0.61	0.54	0.81	0.68	0.58	0.74	0.69	0.64

four systems (AML, Lily, LogMap, Wiktionary) have a wide variation of confidence values.

When comparing the performance of the matchers on the uncertain reference alignments versus that on the sharp version (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 almost to 0). Changes in F-measure of discrete cases ranged from -1 to 17 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 the matches they find are the ones that the experts had high agreement about, while the ones they missed were the more controversial matches. 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, three (DOME, LogMap, SANOM) have relatively small drops in F-measure when moving from discrete to continuous evaluation. Lily's performance drops drastically under the discrete and continuous evaluation methodologies. This is because the matcher assigns low confidence values to some matches in which the labels are equivalent strings, which many crowdsourcers agreed with unless there was a compelling technical reason not to. This hurts recall significantly.

Overall, in comparison with last year, the F-measures of most returning matching systems essentially held constant when evaluated against the uncertain reference alignments. The exception was Lily, whose performance in the discrete case decreased dramatically. ONTMAT1 and Wiktionary are two new systems participating in this year. ONTMAT1's performance in both discrete and continuous cases increases 16 percent in terms of F-measure over the sharp reference alignment from 0.55 to 0.64, which it is mainly driven by increased recall. Wiktionary assigns confidence value of 1.0 to the entities with identical strings in two ontologies, while gives confidence value of 0.5 to other possible candidates. From the results, its performance improves significantly from sharp to discrete and continuous cases.

4.5 Disease and Phenotype Track

In the OAEI 2019 phenotype track 8 systems were able to complete at least one of the tasks with a 6 hours timeout. Table 11 shows 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 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

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	43	2,130	1	0.88	0.85	0.82	0	0.0%
LogMapBio	1,740	2,201	50	0.86	0.85	0.83	0	0.0%
AML	90	2,029	330	0.89	0.84	0.80	0	0.0%
LogMapLt	6	1,370	2	1.00	0.75	0.60	0	0.0%
POMAP++	1,862	1,502	218	0.86	0.68	0.57	0	0.0%
FCAMapKG	14	734	0	1.00	0.49	0.32	0	0.0%
DOME	11	692	0	1.00	0.47	0.30	0	0.0%
Wiktionary	745	61,872	60,634	0.02	0.04	0.55	0	0.0%
DOID-ORDO task								
LogMapBio	2,312	2,547	123	0.91	0.86	0.81	0	0.0%
LogMap	24	2,323	0	0.95	0.85	0.77	0	0.0%
POMAP++	2,497	2,563	192	0.89	0.84	0.79	0	0.0%
LogMapLt	8	1,747	20	0.99	0.75	0.60	0	0.0%
AML	173	4,781	2,342	0.52	0.65	0.87	0	0.0%
FCAMapKG	23	1,274	2	1.00	0.61	0.44	0	0.0%
DOME	17	1,235	5	0.99	0.60	0.43	0	0.0%
Wiktionary	531	909	366	0.57	0.28	0.18	7	0.067%

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 and LogMapBio are the systems that provide the closest set of correspondences to the consensus (not necessarily the best system) in both tasks. LogMap has a small set of unique correspondences as most of its correspondences are also suggested by its variant LogMapBio and vice versa. By contrast, AML and Wiktionary produce the highest number of unique correspondences in HP-MP and DOID-ORDO respectively, and the second-highest inversely. Nonetheless, Wiktionary suggests a very large number of correspondences with respect to the other systems which suggest that it may also include many subsumption and related correspondences and not only equivalence. All systems produce coherent alignments except for Wiktionary in the DOID-ORDO task.

4.6 Large Biomedical Ontologies

In the OAEI 2019 Large Biomedical Ontologies track, 10 systems were able to complete at least one of the tasks within a 6 hours timeout. Eight systems were able to complete all six tasks.[16] The evaluation results for the largest matching tasks are shown in Table 12.

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

[16] 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	75	3,110	276	0.81	0.84	0.88	4	0.012%
LogMap	82	2,701	0	0.86	0.83	0.81	3	0.009%
LogMapBio	2,072	3,104	139	0.78	0.81	0.85	3	0.009%
LogMapLt	9	3,458	75	0.68	0.74	0.82	8,925	27.3%
Wiktionary	4,699	1,873	56	0.93	0.73	0.61	3,476	10.6%
DOME	21	2,413	7	0.80	0.73	0.67	1,033	3.2%
FCAMapKG	0	3,765	316	0.62	0.71	0.82	10,708	32.8%
AGM	3,325	7,648	6,819	0.08	0.12	0.22	28,537	87.4%
Whole FMA ontology with SNOMED large fragment (Task 4)								
LogMap	394	6,393	0	0.84	0.73	0.65	0	0.0%
LogMapBio	2,853	6,926	280	0.79	0.72	0.67	0	0.0%
AML	152	8,163	2,525	0.69	0.70	0.71	0	0.0%
FCAMapKG	0	1,863	77	0.88	0.36	0.22	1,527	2.0%
LogMapLt	15	1,820	47	0.85	0.33	0.21	1,386	1.8%
DOME	38	1,589	1	0.94	0.33	0.20	1,348	1.8%
Wiktionary	12,633	1,486	143	0.82	0.28	0.17	790	1.0%
AGM	4,227	11,896	10,644	0.07	0.09	0.13	70,923	92.7%
Whole NCI ontology with SNOMED large fragment (Task 6)								
AML	331	14,200	2,656	0.86	0.77	0.69	$\geq$578	$\geq$0.5%
LogMapBio	4,586	13,732	940	0.81	0.71	0.63	$\geq$1	$\geq$0.001%
LogMap	590	12,276	0	0.87	0.71	0.60	$\geq$1	$\geq$0.001%
LogMapLt	16	12,864	658	0.80	0.66	0.57	$\geq$91,207	$\geq$84.7%
FCAMapKG	0	12,813	1,115	0.79	0.65	0.56	$\geq$84,579	$\geq$78.5%
DOME	38	9,806	26	0.91	0.64	0.49	$\geq$66,317	$\geq$61.6%
Wiktionary	9,208	9,585	518	0.90	0.62	0.47	$\geq$65,968	$\geq$61.2%
AGM	5,016	21,600	16,253	0.23	0.25	0.28	-	-

Interestingly, the use of background knowledge led to an improvement in recall from LogMapBio 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.[17] 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 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.

[17] http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2019/results/

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

With respect to alignment coherence, as in previous OAEI editions, only two distinct systems have shown alignment repair facilities: AML, LogMap and its LogMapBio variant. Note that only LogMap and LogMapBio are able to reduce to a minimum the number of unsatisfiable classes across all tasks, missing 3 unsatisfiable classes in the worst case (whole FMA-NCI task). For the AGM correspondences the ELK reasoner could not complete the classification over the integrated ontology within the allocated time.

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 [39], the repair module of LogMap (LogMap-Repair) [31] or the repair module of AML [43], which have worked well in practice [33, 24].

4.7 Multifarm

This year, 5 systems registered to participate in the MultiFarm track: AML, EVOCROS, Lily, LogMap and Wiktionary. This number slightly decreases with respect to the last campaign (6 in 2018, 8 in 2017, 7 in 2016, 5 in 2015, 3 in 2014, 7 in 2013, and 7 in 2012). The reader can refer to the OAEI papers for a detailed description of the strategies adopted by each system. In fact, most systems still adopt a translation step before the matching itself. However, a few systems had issues when evaluated: i) EVOCROS encountered problems to complete a single matching task; and ii) Lily has generated mostly empty alignments.

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 haven't applied any threshold on the results. 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).

AML outperforms all other systems in terms of F-measure for task i) (same behaviour than last year). In terms of precision, the systems have relatively similar results. With respect to the task ii) LogMap has the best performance. AML and LogMap have participated last year. Comparing the results from last year, in terms F-measure (cases of type i), AML maintains its overall performance (.45 in 2019, .46 in 2018, .46 in 2017, .45 in 2016 and .47 in 2015). The same could be observed for LogMap (.37 in 2018, .36 in 2017, and .37 in 2016).

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 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

Table 13. MultiFarm aggregated results per matcher, for each type of matching task – different ontologies (i) and same ontologies (ii). 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.

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	236	55	8.18	.72 (.72)	.45 (.45)	.34 (.34)	33.40	.93 (.95)	.27 (.28)	.17 (.16)
LogMap	49	55	6.99	.72 (.72)	.37 (.37)	.25 (.25)	46.80	.95 (.96)	.41 (.42)	.28 (.28)
Wiktionary	785	23	4.91	.76 (.79)	.31 (.33)	.21 (.22)	9.24	.94 (.96)	.12 (.12)	.07 (.06)

in Wikipedia, BabelNet, etc.) while strategies such as machine learning, or indirect alignment composition remain under-exploited.

4.8 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. Those were the exact same systems (and versions) that participated on OAEI 2018.

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 9.7s and 360s, respectively, to complete the two tasks. The results can also be found in HOBBIT platform (`https://tinyurl.com/yywwlsmt` - Login as Guest).

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. The results can also be found in HOBBIT platform (`https://tinyurl.com/y4vk6htq` - Login as Guest).

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*.

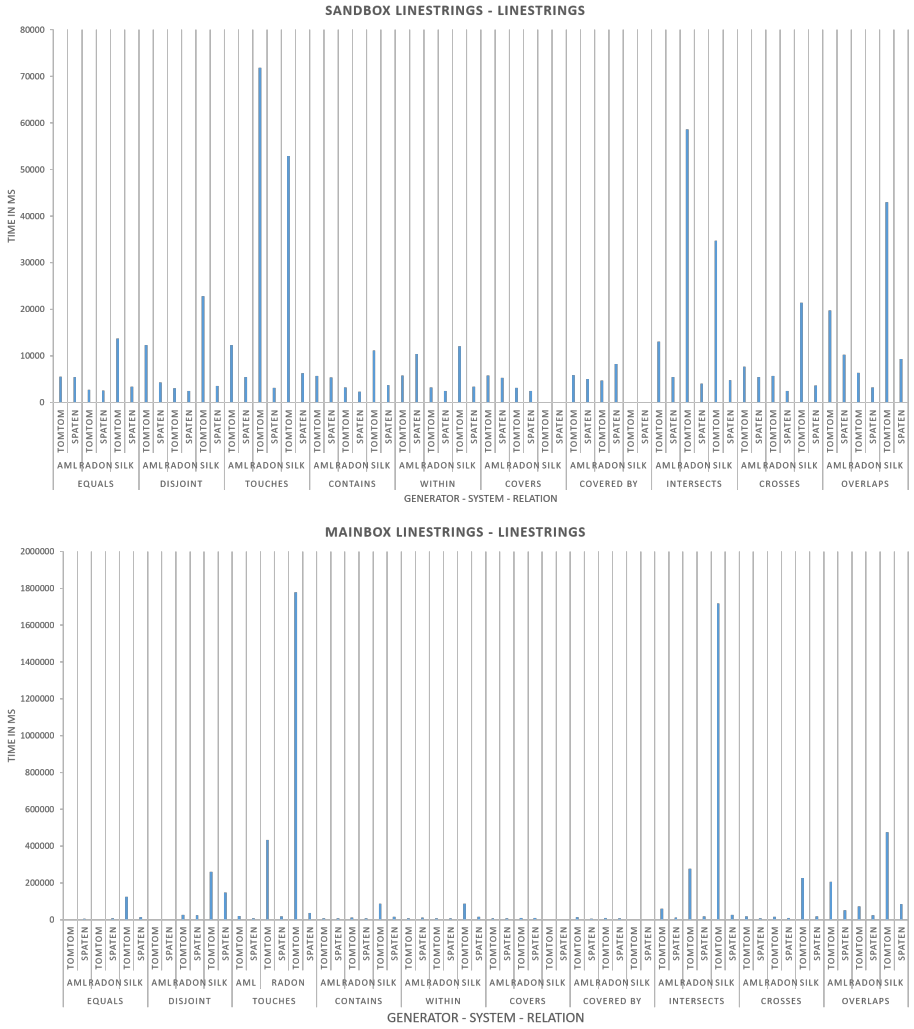

Fig. 2. Time performance for TomTom & Spaten SLL (top) and LLL (bottom) suites for AML (A), Silk (S) and RADON (R).

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.

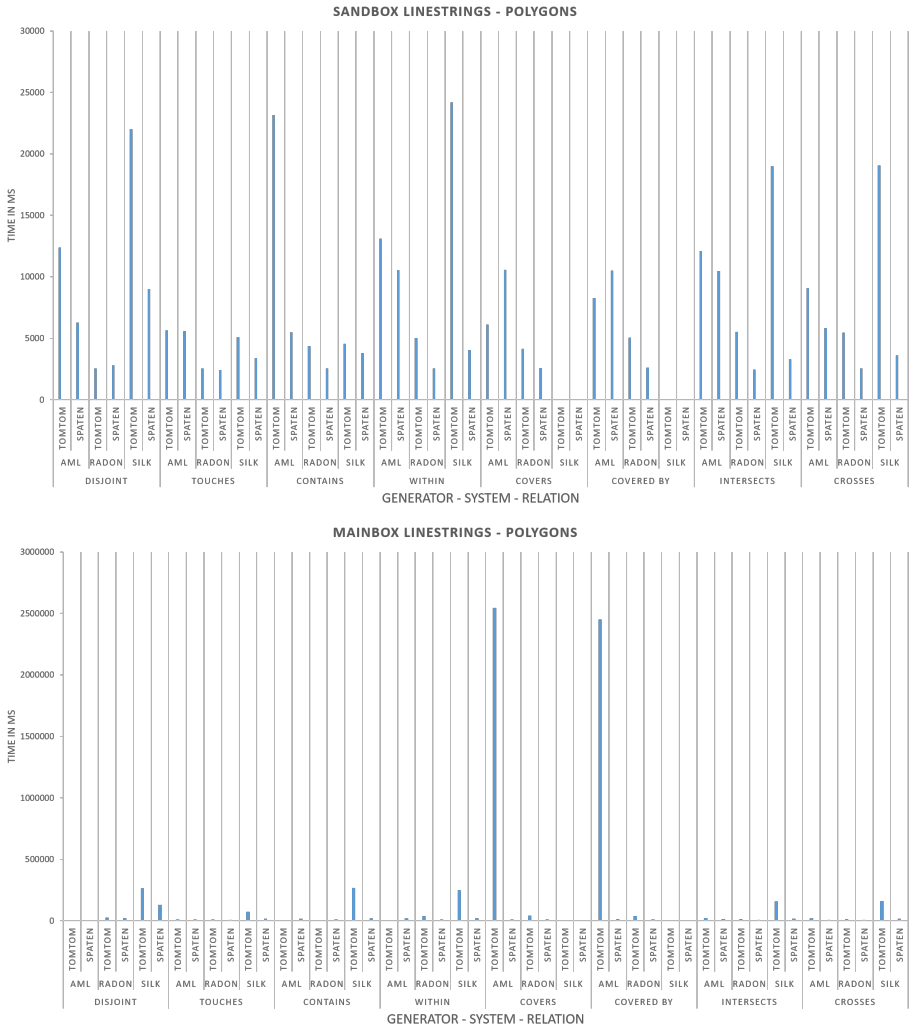

Fig. 3. Time performance for TomTom & Spaten SLP (top) and LLP (bottom) suites for AML (A), Silk (S) and RADON (R).

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.

4.9 SPIMBENCH

This year, the SPIMBENCH track counted four participants: AML, Lily, LogMap and FTRLIM. FTRLIM participated for the first time this year while AML, Lily, and LogMap also participated the previous years. The evaluation results of the track are shown in Table 14. The results can also be found in HOBBIT platform (`https://tinyurl.com/yxhsw48c` - Login as Guest).

Table 14. SPIMBENCH track results.

System	Precision	Recall	F-measure	Time (ms)
Sandbox (100 instances)				
AML	0.8348	0.8963	0.8645	6223
Lily	0.8494	1.0	0.9185	2032
LogMap	0.9382	0.7625	0.8413	6919
FTRLIM	0.8542	1.0	0.9214	1474
Mainbox (5000 instances)				
AML	0.8385	0.8835	0.8604	39515
Lily	0.8546	1.0	0.9216	3667
LogMap	0.8925	0.7094	0.7905	26920
FTRLIM	0.8558	1.0	0.9214	2155

Lily and FTRLIM had the best performance overall both in terms of F-measure and run time. Notably, their run time scaled very well with the increase in the number of instances. Lily, FTRLIM, and AML had a higher recall than precision, while Lily and FTRLIM had a full recall. By contrast, LogMap had the highest precision but lowest

recall of all the 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.

4.10 Knowledge Graph

We evaluated all SEALS participants in the OAEI (even those not registered for the track) on a very small matching task[18]. This revealed that not all systems were able to handle the task, and in the end, only the following systems were evaluated: AGM, AML, DOME, FCAMap-KG, LogMap, LogMapBio, LogMapKG, LogMapLt, POMap++, Wiktionary. Out of those only LogMapBio, LogMapLt and POMap++ were not registered for this track. In comparison to last year, more matchers participate and return meaningful correspondences. Moreover there are systems which especially focus on the knowledge graph track e.g. FCAMap-KG and LogMapKG.

Table 15 shows the aggregated results for all systems, including the number of tasks in which they were 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 which are shown between parentheses in the table, where applicable.

Nearly all systems were able to generate class correspondences. In terms of F-measure, AML is the best one (when considering only completed test cases). Many matchers were also able to beat the baseline. The highest recall is about 0.77 which shows that some class correspondences are not easy to find.

In comparison to last year, more matchers are able to produce property correspondences. Only the systems of the LogMap family and POMAP++ do not return any alignments. While Wiktionary and FCAMap-KG achieve an F-Measure of 0.98, other systems need more improvement here because they are not capable of beating the baseline (mostly due to low recall).

With respect to instance correspondences, AML and DOME are the best systems, but they outperform the baselines only by a small margin. On average, the systems returned between 3,000 and 8,000 instance alignments. Only LogMapKG returned nearly 30,000 mappings. This is interesting because it should be focused on generating only 1:1 alignments, but deviates here.

We also analyzed the arity of the resulting alignments because in the knowledge graph track it is probably better to focus on a 1:1 mapping. Such a strict mapping is returned by the following systems: AGM, baselineLabel, DOME and POMAP++. LogMap and LogMapBio return a few correspondences with same source or target in only two test cases. BaselineAltLabel, FCAMap-KG and Wiktionary returned some n:m mappings in all test cases. AML and LogMapLt returned more of those and LogMapKG has the highest amount of n:m mappings.

When analyzing the confidence values of the alignments, it turns out that most matchers set it to 1 (AGM,baselineAltLabel, baselineLabel, FCAMap-KG, LogMapLt,

[18] `http://oaei.ontologymatching.org/2019/results/knowledgegraph/small_test.zip`

Table 15. Knowledge Graph track results, divided into class, property, instance, and overall correspondences.

System	Time (s)	# tasks	Size	Prec.	F-m.	Rec.
Class performance						
AGM	10:47:38	5	14.6	0.23	0.09	0.06)
AML	0:45:46	4	27.5	0.78 (0.98)	0.69 (0.86)	0.61 (0.77)
baselineAltLabel	0:11:48	5	16.4	1.0	0.74	0.59
baselineLabel	0:12:30	5	16.4	1.0	0.74	0.59
DOME	1:05:26	4	22.5	0.74 (0.92)	0.62 (0.77)	0.53 (0.66)
FCAMap-KG	1:14:49	5	18.6	1.0	0.82	0.70
LogMap	0:15:43	5	26.0	0.95	0.84	0.76)
LogMapBio	2:31:01	5	26.0	0.95	0.84	0.76)
LogMapKG	2:26:14	5	26.0	0.95	0.84	0.76)
LogMapLt	0:07:28	4	23.0	0.80 (1.0)	0.56 (0.70)	0.43 (0.54)
POMAP++	0:14:39	5	2.0	0.0	0.0	0.0
Wiktionary	0:20:14	5	21.4	1.0	0.8	0.67
Property performance						
AGM	10:47:38	5	49.4	0.66	0.32	0.21)
AML	0:45:46	4	58.2	0.72 (0.91)	0.59 (0.73)	0.49 (0.62)
baselineAltLabel	0:11:48	5	47.8	0.99	0.79	0.66
baselineLabel	0:12:30	5	47.8	0.99	0.79	0.66
DOME	1:05:26	4	75.5	0.79 (0.99)	0.77 (0.96)	0.75 (0.93)
FCAMap-KG	1:14:49	5	69.0	1.0	0.98	0.96
LogMap	0:15:43	5	0.0	0.0	0.0	0.0)
LogMapBio	2:31:01	5	0.0	0.0	0.0	0.0)
LogMapKG	2:26:14	5	0.0	0.0	0.0	0.0)
LogMapLt	0:07:28	4	0.0	0.0	0.0	0.0)
POMAP++	0:14:39	5	0.0	0.0	0.0	0.0)
Wiktionary	0:20:14	5	75.8	0.97	0.98	0.98
Instance performance						
AGM	10:47:38	5	5169.0	0.48	0.25	0.17)
AML	0:45:46	4	7529.8	0.72 (0.90)	0.71 (0.88)	0.69 (0.86)
baselineAltLabel	0:11:48	5	4674.2	0.89	0.84	0.80
baselineLabel	0:12:30	5	3641.2	0.95	0.81	0.71
DOME	1:05:26	4	4895.2	0.74 (0.92)	0.70 (0.88)	0.67 (0.84)
FCAMap-KG	1:14:49	5	4530.6	0.90	0.84	0.79
LogMap	0:15:43	5	0.0	0.0	0.0	0.0)
LogMapBio	2:31:01	5	0.0	0.0	0.0	0.0)
LogMapKG	2:26:14	5	29190.4	0.40	0.54	0.86)
LogMapLt	0:07:28	4	6653.8	0.73 (0.91)	0.67 (0.84)	0.62 (0.78)
POMAP++	0:14:39	5	0.0	0.0	0.0	0.0
Wiktionary	0:20:14	5	3483.6	0.91	0.79	0.70
Overall performance						
AGM	10:47:38	5	5233.2	0.48	0.25	0.17)
AML	0:45:46	4	7615.5	0.72 (0.90)	0.70 (0.88)	0.69 (0.86)
baselineAltLabel	0:11:48	5	4739.0	0.89	0.84	0.80
baselineLabel	0:12:30	5	3706.0	0.95	0.81	0.71
DOME	1:05:26	4	4994.8	0.74 (0.92)	0.70 (0.88)	0.67 (0.84)
FCAMap-KG	1:14:49	5	4792.6	0.91	0.85	0.79
LogMap	0:15:43	5	26.0	0.95	0.01	0.0)
LogMapBio	2:31:01	5	26.0	0.95	0.01	0.0)
LogMapKG	2:26:14	5	29216.4	0.40	0.54	0.84)
LogMapLt	0:07:28	4	6676.8	0.73 (0.91)	0.66 (0.83)	0.61 (0.76)
POMAP++	0:14:39	5	19.4	0.0	0.0	0.0
Wiktionary	0:20:14	5	3581.8	0.91	0.8	0.71

Wiktionary). AML and LogMapKG set it higher than 0.6 whereas only DOME uses the full range between zero and one. LogMap and LogMapBio uses a range of 0.3 and 0.8. The confidences were analyzed with the MELT dashboard[19] [28].

Regarding runtime, AGM (10:47:38) was the slowest system, followed by LogMapKG and LogMapBio which were much faster. Besides AGM all five test cases could be completed in under 3 hours.

4.11 Interactive matching

This year, three systems participated in the Interactive matching track. They are ALIN, AML, and LogMap. Their results are shown in Table 16 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.

The matching systems that participated in this track employ different user-interaction strategies. While LogMap, 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

[19] http://oaei.ontologymatching.org/2019/results/knowledgegraph/knowledge_graph_dashboard.html

Table 16. 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	0.974	0.698	0.813	0.365	–	–	–	–	–	–	–
	0.0	0.979	0.85	0.91	0.63	0.979	0.85	0.91	365	638	1.0	1.0
	0.1	0.953	0.832	0.889	0.599	0.979	0.848	0.909	339	564	0.854	0.933
	0.2	0.929	0.817	0.869	0.569	0.979	0.848	0.909	332	549	0.728	0.852
	0.3	0.908	0.799	0.85	0.54	0.979	0.847	0.908	326	536	0.616	0.765
AML	NI	0.95	0.936	0.943	0.832	–	–	–	–	–	–	–
	0.0	0.968	0.948	0.958	0.862	0.968	0.948	0.958	236	235	1.0	1.0
	0.1	0.954	0.944	0.949	0.853	0.969	0.947	0.958	237	235	0.696	0.973
	0.2	0.944	0.94	0.942	0.846	0.969	0.948	0.959	252	248	0.565	0.933
	0.3	0.935	0.933	0.933	0.827	0.969	0.946	0.957	238	234	0.415	0.878
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	0.962	0.831	0.892	0.566	0.964	0.803	0.876	388	1164	0.752	0.965
	0.2	0.945	0.822	0.879	0.549	0.945	0.763	0.844	388	1164	0.57	0.926
	0.3	0.933	0.815	0.87	0.535	0.921	0.724	0.811	388	1164	0.432	0.872
Conference Dataset												
ALIN	NI	0.871	0.443	0.587	–	–	–	–	–	–	–	–
	0.0	0.914	0.695	0.79	–	0.914	0.695	0.79	228	373	1.0	1.0
	0.1	0.809	0.658	0.725	–	0.919	0.704	0.798	226	367	0.707	0.971
	0.2	0.715	0.631	0.67	–	0.926	0.717	0.808	221	357	0.5	0.942
	0.3	0.636	0.605	0.62	–	0.931	0.73	0.819	219	353	0.366	0.908
AML	NI	0.841	0.659	0.739		–	–	–	–	–	–	–
	0.0	0.91	0.698	0.79	–	0.91	0.698	0.79	221	220	1.0	1.0
	0.1	0.846	0.687	0.758	–	0.916	0.716	0.804	242	236	0.726	0.971
	0.2	0.783	0.67	0.721	–	0.924	0.729	0.815	263	251	0.571	0.933
	0.3	0.721	0.646	0.681	–	0.927	0.741	0.824	273	257	0.446	0.877
LogMap	NI	0.818	0.59	0.686	–	–	–	–	–	–	–	–
	0.0	0.886	0.61	0.723	–	0.886	0.61	0.723	82	246	1.0	1.0
	0.1	0.845	0.595	0.698	–	0.857	0.576	0.689	82	246	0.694	0.973
	0.2	0.818	0.586	0.683	–	0.827	0.546	0.657	82	246	0.507	0.941
	0.3	0.799	0.588	0.677	–	0.81	0.519	0.633	82	246	0.376	0.914

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

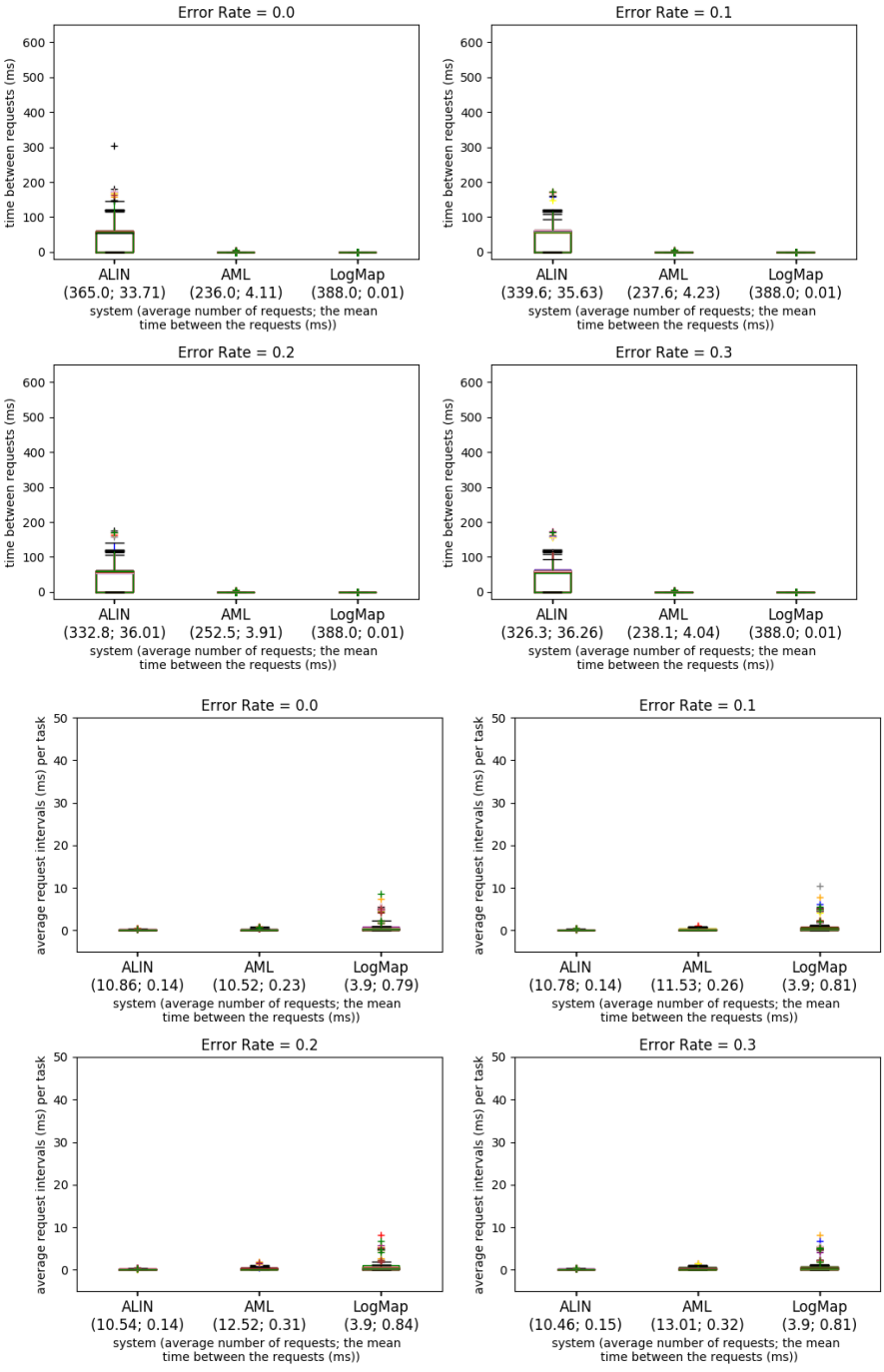

Fig. 4. Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

patterns or their involvement in unsatisfiabilities) and AML presents one correspondence at a time to the user. 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. 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, and AML 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 [11]: 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).

4.12 Complex Matching

Three systems were able to generate complex correspondences: AMLC, AROA [53], and CANARD. The results for the other systems are reported in terms of simple alignments. The results of the systems on the five test cases are summarized in Table 17.

With respect to the Hydrography test case, only AMLC can generate two correct complex correspondences which are stating that a class in the source ontology is equivalent to the union of two classes in the target ontology. Most of the systems achieved fair results in terms of precision, but the low recall reflects that the current ontology alignment systems still need to be improved to find more complex relations.

In terms of GeoLink test cases, the real-world instance data from GeoLink Project is also populated into the ontology in order to enable the systems that depend on instance-based matching algorithms to evaluate their performance. There are three alignment

Table 17. Results of the Complex Track in OAEI 2019.

Matcher	Conference			Populated Conference		Hydrography			GeoLink			Taxon	
	Prec.	F-meas.	Rec.	Prec.	Coverage	relaxed_Prec.	relaxed_F-meas.	relaxed_Rec.	relaxed_Prec.	relaxed_F-meas.	relaxed_Rec.	Prec.	Coverage
AGM	-	-	-	-	-	-	-	-	-	-	-	0.06 - 0.14	0.03 - 0.04
Alin	-	-	-	**0.68 - 0.98**	0.20 - 0.28	-	-	-	-	-	-	-	-
AML	-	-	-	0.59 - 0.93	0.31 - 0.37	-	-	-	-	-	-	0.53	0.00
AMLC	**0.31**	**0.34**	**0.37**	0.30 - 0.59	**0.46 - 0.50**	0.45	0.10	0.05	0.50	0.32	0.23	-	-
AROA	-	-	-	-	-	-	-	-	0.87	**0.60**	**0.46**	-	-
CANARD	-	-	-	0.21 - 0.88	0.40 - **0.51**	-	-	-	0.89	0.54	0.39	0.08 - 0.91	**0.14 - 0.36**
DOME	-	-	-	0.59 - 0.94	0.40 - 0.51	-	-	-	-	-	-	-	-
FcaMapKG	-	-	-	0.51 - 0.82	0.21 - 0.28	-	-	-	-	-	-	0.63 - 0.96	0.03 - 0.05
Lily	-	-	-	0.45 - 0.73	0.23 - 0.28	-	-	-	-	-	-	-	-
LogMap	-	-	-	**0.56 - 0.96**	0.25 - 0.32	0.67	0.10	0.05	0.85	0.29	0.18	0.63 - 0.79	0.11 - 0.14
LogMapBio	-	-	-	-	-	**0.70**	0.10	0.05	-	-	-	0.54 - 0.72	0.08 - 0.11
LogMapKG	-	-	-	**0.56 - 0.96**	0.25 - 0.32	0.67	0.10	0.05	-	-	-	0.55 - 0.69	**0.14** - 0.17
LogMapLt	-	-	-	0.50 - 0.87	0.23 - 0.32	0.67	0.10	0.05	-	-	-	0.54 - 0.72	0.08 - 0.11
ONTMAT1	-	-	-	**0.67 - 0.98**	0.20 - 0.28	-	-	-	-	-	-	-	-
POMAP++	-	-	-	0.25 - 0.54	0.20 - 0.29	0.65	0.07	0.04	**0.90**	0.26	0.16	1.00	0.00
Wikitionary	-	-	-	0.48 - 0.88	0.26 - 0.34	-	-	-	-	-	-	-	-

systems that generate complex alignments in GeoLink Benchmark, which are AMLC, AROA, and CANARD. AMLC didn't find any correct complex alignment, while AROA and CARARD achieved relatively good performance. One of the reasons may be that these two systems are instance-based systems, which rely on the shared instances between ontologies. In other words, the shared instance data between two ontologies would be helpful to the matching process.

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

With respect to the Conference test cases although the performance in terms of precision and recall decreased for AMLC, AMLC managed to find more true positives than the last year. Since AMLC provides confidence, it could be possible to include confidence into the evaluation and this could improve the performance results. AMLC discovered one more kind of complex mappings: the union of classes.

A more detailed discussion of the results of each task can be found in the OAEI page for this track. For a second 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

In 2019, 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., [30]).

According to the *Conference* track there is still need for an improvement with regard to the ability of matching systems to match properties. To assist system developers in tackling this aspect we provided a more detailed evaluation in terms of the analysis of the false positives per matching system (available on the Conference track web page). However, this could be extended by the inspection of the reasons why the matching system found the given false positives.As already pointed out last year, 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.

Despite the quite promising results obtained by matching systems for the **Biodiversity and Ecology track**, the most important observation is that none of the systems has been able to detect mappings established by the experts. Detecting such correspondences requires the use of domain-specific core knowledge that captures biodiversity concepts. We expect this domain-specific background to be integrated in future versions of the systems.

The **interactive matching track** also witnessed a small number of participants. Three systems participated this year. This is puzzling considering that this track is based on the *Anatomy* and *Conference* test cases, and those tracks had 13 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.

The **complex matching track** opens new perspectives in the field of ontology matching. Tackling complex matching automatically is extremely challenging, likely requiring profound adaptations from matching systems, so the fact that there were three participants that were 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. This year automatic evaluation has been introduced following an instance-based comparison approach.

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 *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.

In the **knowledge graph track**, we could observe that simple baselines are still hard to beat – which was also the case in other tracks when they were still new. We expect more sophisticated and powerful implementations in the next 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. 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 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 CNRS Blanc project RegleX-LD.

Daniel Faria was supported by the EC H2020 grant 676559 ELIXIR-EXCELERATE and the Portuguese FCT Grant 22231 BioData.pt, co-financed by FEDER.

Ernesto Jimenez-Ruiz has been partially supported by the SIRIUS Centre for Scalable Data Access (Research Council of Norway, project no.: 237889) and the AIDA project (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 EU's Horizon 2020 research and innovation programme under grant agreement No 688227 (Hobbit).

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

Patrick Lambrix and Huanyu Li have been supported by the Swedish e-Science Research Centre (SeRC), the Swedish Research Council (Vetenskapsrådet, dnr 2018-04147) and the Swedish National Graduate School in Computer Science (CUGS).

The Biodiversity and Ecology track has been partially funded by the German Research Foundation in the context of the GFBio Project (grant No. SE 553/7-1) and the CRC 1076 AquaDiva, the Leitprojekt der Fraunhofer Gesellschaft in the context of the MED2ICIN project (grant No. 600628) and the German Network for Bioinformatics Infrastructure - de.NBI (grant No. 031A539B).

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*, 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)*, 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. Alsayed Algergawy, Michelle Cheatham, Daniel Faria, Alfio Ferrara, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Naouel Karam, Abderrahmane Khiat, Patrick Lambrix, Huanyu Li, Stefano Montanelli, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Daniela Schmidt, Pavel Shvaiko, Andrea Splendiani, Élodie Thiéblin, Cássia Trojahn, Jana Vatascinová, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2018. In *Proceedings of the 13th International Workshop on Ontology Matching, Monterey (CA, US)*, pages 76–116, 2018.
5. Benhamin Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Proc. K-Cap Workshop on Integrating Ontologies*, Banff (Canada), 2005.
6. Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.
7. Caterina Caracciolo, Jérôme Euzenat, Laura Hollink, Ryutaro Ichise, Antoine Isaac, Véronique Malaisé, Christian Meilicke, Juan Pane, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, and Vojtech Svátek. Results of the ontology alignment evaluation initiative 2008. In *Proceedings of the 3rd Ontology matching workshop, Karlsruhe (DE)*, pages 73–120, 2008.
8. Michelle Cheatham, Zlatan Dragisic, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, 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.
9. Michelle Cheatham, Dalia Varanka, Fatima Arauz, and Lu Zhou. Alignment of surface water ontologies: a comparison of manual and automated approaches. *Journal of Geographical Systems*, pages 1–23, 2019.
10. 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.
11. 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.

12. 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.
13. Zlatan Dragisic, Kai Eckert, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, Andreas Oskar Kempf, Patrick Lambrix, Stefano Montanelli, Heiko Paulheim, Dominique Ritze, Pavel Shvaiko, Alessandro Solimando, Cássia Trojahn dos Santos, Ondrej Zamazal, and Bernardo Cuenca Grau. Results of the ontology alignment evaluation initiative 2014. In *Proceedings of the 9th International Ontology matching workshop, Riva del Garda (IT)*, pages 61–104, 2014.
14. Zlatan Dragisic, Valentina Ivanova, Patrick Lambrix, Daniel Faria, Ernesto Jiménez-Ruiz, and Catia Pesquita. User validation in ontology alignment. In *Proceedings of the 15th International Semantic Web Conference, Kobe (JP)*, pages 200–217, 2016.
15. Zlatan Dragisic, Valentina Ivanova, Huanyu Li, and Patrick Lambrix. Experiences from the anatomy track in the ontology alignment evaluation initiative. *Journal of Biomedical Semantics*, 8:56:1–56:28, 2017.
16. Marc Ehrig and Jérôme Euzenat. Relaxed precision and recall for ontology matching. In *Integrating Ontologies, Proceedings of the K-CAP Workshop on Integrating Ontologies, Banff, Canada*, 2005.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, 2nd edition, 2013.
24. 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, 2014.

25. 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*, 8:55:1–55:13, 2017.
26. Sven Hertling and Heiko Paulheim. Dbkwik: A consolidated knowledge graph from thousands of wikis. In *Proceedings of the International Conference on Big Knowledge*, 2018.
27. Sven Hertling and Heiko Paulheim. Dbkwik: extracting and integrating knowledge from thousands of wikis. *Knowledge and Information Systems*, 2019.
28. Sven Hertling, Jan Portisch, and Heiko Paulheim. Melt - matching evaluation toolkit. In *SEMANTICS*, 2019.
29. 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.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. Friederike Klan, Erik Faessler, Alsayed Algergawy, Birgitta König-Ries, and Udo Hahn. Integrated semantic search on structured and unstructured data in the adonis system. In *Proceedings of the 2nd International Workshop on Semantics for Biodiversity*, 2017.
38. Huanyu Li, Zlatan Dragisic, Daniel Faria, Valentina Ivanova, Ernesto Jiménez-Ruiz, Patrick Lambrix, and Catia Pesquita. User validation in ontology alignment: functional assessment and impact. *The Knowledge Engineering Review*, 34:e15, 2019.
39. Christian Meilicke. *Alignment Incoherence in Ontology Matching*. PhD thesis, University Mannheim, 2011.
40. 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.

41. Boris Motik, Rob Shearer, and Ian Horrocks. Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009.
42. 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.
43. 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.
44. 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.
45. 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.
46. Alessandro Solimando, Ernesto Jimenez-Ruiz, and Giovanna Guerrini. Minimizing conservativity violations in ontology alignments: Algorithms and evaluation. *Knowledge and Information Systems*, 2016.
47. Christian Strobl. *Encyclopedia of GIS*, chapter Dimensionally Extended Nine-Intersection Model (DE-9IM), pages 240–245. Springer, 2008.
48. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the Workshop on Evaluation of Ontology-based Tools (EON), Hiroshima (JP)*, 2004.
49. Élodie Thiéblin. Do competency questions for alignment help fostering complex correspondences? In *Proceedings of the EKAW Doctoral Consortium 2018*, 2018.
50. Élodie Thiéblin, Fabien Amarger, Ollivier Haemmerlé, Nathalie Hernandez, and Cássia Trojahn dos Santos. Rewriting SELECT SPARQL queries from 1: n complex correspondences. In *Proceedings of the 11th International Workshop on Ontology Matching*, pages 49–60, 2016.
51. 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.
52. 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.
53. Lu Zhou, Michelle Cheatham, and Pascal Hitzler. Towards association rule-based complex ontology alignment. In *Proceedings of the 9th Joint International Semantic Technology Conference JIST 2019, Hangzhou, China, November 25*, in press, 2019.
54. 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)*, pages 273–288, 2018.
55. Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. Geolink dataset: a complex alignment benchmark from real-world ontology. *Data Intelligence*, in press, 2019.

Jena, Lisboa, Milano, Heraklion, Mannheim,
Oslo, London, Berlin, Bonn, Linköping,
Trento, Toulouse, Prague, Manhattan
November 2019