Results of the Ontology Alignment Evaluation Initiative 2020

Mina Abd Nikooie Pour, Alsayed Algergawy, Reihaneh Amini, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Clement Jonquet, Naouel Karam, et al.

Results of the Ontology Alignment Evaluation Initiative 2020*

Mina Abd Nikooie Pour[1], Alsayed Algergawy[2], Reihaneh Amini[3], Daniel Faria[4], Irini Fundulaki[5], Ian Harrow[6], Sven Hertling[7], Ernesto Jiménez-Ruiz[8,9], Clement Jonquet[10], Naouel Karam[11], Abderrahmane Khiat[12], Amir Laadhar[10], Patrick Lambrix[1], Huanyu Li[1], Ying Li[1], Pascal Hitzler[3], Heiko Paulheim[7], Catia Pesquita[13], Tzanina Saveta[5], Pavel Shvaiko[14], Andrea Splendiani[6], Elodie Thiéblin[15], Cássia Trojahn[16], Jana Vataščinová[17], Beyza Yaman[18], Ondřej Zamazal[17], and Lu Zhou[3]

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

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 and use different evaluation modalities (e.g., blind evaluation, open evaluation, or consensus). The OAEI 2020 campaign offered 12 tracks with 36 test cases, and was attended by 19 participants. This paper is an overall presentation of that campaign.

1 Introduction

The Ontology Alignment Evaluation Initiative[1] (OAEI) is a coordinated international initiative, which organizes the evaluation of an increasing number of ontology matching systems [26, 28], and which has been run for seventeen years by now. 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, the ambition is that, from such evaluations, developers can improve their systems and offer better tools that answer the evolving application needs.

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) [66]. 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) [7]. From 2006 until the present, the OAEI campaigns were held at the Ontology Matching workshop, collocated with ISWC [5, 4, 1, 2, 11, 18, 15, 3, 24, 23, 22, 10, 25, 27], which this year took place virtually (originally planned in Athens, Greece)[2].

Since 2011, we have been using an environment for automatically processing evaluations (Section 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 (Section 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 year, the MELT framework [36] was adopted in order to facilitate the SEALS and HOBBIT wrapping and evaluation.

This paper synthesizes the 2020 evaluation campaign and introduces the results provided in the papers of the participants. The remainder of the paper is organized as follows: in Section 2, we present the overall evaluation methodology; in Section 3 we present the tracks and datasets; in Section 4 we present and discuss the results; and finally, Section 5 discusses the lessons learned.

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

[2] http://om2020.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 [44].

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.

The **MELT framework**[5] [36] was introduced in 2019 and is under active development. It allows to develop, evaluate, and package matching systems for arbitrary evaluation interfaces like SEALS or HOBBIT. It further enables developers to use Python in their matching systems. In terms of evaluation, MELT offers a correspondence level analysis for multiple matching systems which can even implement different interfaces. It is, therefore, suitable for track organisers as well as system developers.

2.2 OAEI campaign phases

As in previous years, the OAEI 2020 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}, 2020. 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 October 15^{th}, 2020, at which date participants had to submit the (near) final versions of their systems (SEALS-wrapped and/or HOBBIT-wrapped).

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

[5] `https://github.com/dwslab/melt`

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 24^{th}, 2020.

3 Tracks and test cases

This year's OAEI campaign consisted of 12 tracks gathering 36 test cases, all of which included OWL ontologies to align.[6] 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.

3.1 Anatomy

The anatomy track comprises a single test case consisting of matching two fragments of biomedical ontologies which describe the human anatomy[7] (3304 classes) and the anatomy of the mouse[8] (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 [20].

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. For some system requires more RAM, the evaluation was carried out on a Windows 10 (64-bit) desktop with an Intel Core i7-6700 CPU @ 3.40GHz x 8 with 16GB RAM allocated. 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.

[6] The Biodiversity and Ecology track also included SKOS thesauri.

[7] `www.cancer.gov/cancertopics/cancerlibrary/terminologyresources`

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

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	4	=	[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
Geolink Cruise	4	=	[0 1]	open	EN	SEALS
Instance and Schema Matching						
Knowledge Graph	5	=	[0 1]	open+blind	EN	SEALS
Interactive Matching						
Interactive	2 (22)	=, <=	[0 1]	open	EN	SEALS
Complex Matching						
Complex	7	=, <=, >=	[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.

3.2 Biodiversity and Ecology

The biodiversity and ecology (biodiv) track has been originally motivated by two projects, namely GFBio[9] (The German Federation for Biological Data) and AquaDiva[10], which aim at providing semantically enriched data management solutions for data capture, annotation, indexing and search [46, 48]. This year, the third edition of the biodiv track features the two matching tasks present in former editions, namely: matching the Environment Ontology (ENVO) [9] to the Semantic Web for Earth and Environment Technology Ontology (SWEET) [58], and matching the Flora Phenotype Ontology (FLOPO) [38] to Plant Trait Ontology (PTO) [14]. In this edition, we partnered with the D2KAB project[11] (Data to Knowledge in Agronomy and Biodiversity) which develops the AgroPortal[12] vocabulary and ontology repository, to include

[9] www.gfbio.org
[10] www.aquadiva.uni-jena.de
[11] www.d2kab.org
[12] agroportal.lirmm.fr

two new matching tasks involving important thesauri (originally developed in SKOS) in agronomy and environmental sciences: finding alignments between the AGROVOC thesaurus [59] and the US National Agricultural Library Thesaurus (NALT)[13] and between the General Multilingual Environmental Thesaurus (GEMET)[14] and the Analysis and Experimentation on Ecosystems thesaurus (ANAEETHES)[13]. These ontologies and thesauri are particularly useful for biodiversity and ecology research and are being used in various projects. They have been developed in parallel and are significantly overlapping. They are semantically rich and contain tens of thousands of concepts. By providing semantic resources developed in SKOS, our objective is also to encourage the ontology alignment community to develop tools that can natively handle SKOS which is an important standard to encode terminologies (particularly thesauri and taxonomies) and for which alignment is also very important.

Table 2 presents detailed information about the ontologies and thesauri used in the evaluation, such as the ontology format, version, number of classes as well as the number of instances[15].

Table 2. Version, format and number of classes of the Biodiversity and Ecology track ontologies and thesauri.

Ontology/Thesaurus	Format	Version	Classes	Instances
ENVO	OWL	2020-03-08	9053	-
SWEET	OWL	2019-10-12	4533	-
FLOPO	OWL	2016-06-03	28965	-
PTO	OWL	2017-09-11	1504	-
AGROVOC	SKOS	2020-10-02	46	706803
NALT	SKOS	2020-28-01	2	74158
GEMET	SKOS	2020-13-02	7	5907
ANAEETHES	SKOS	2017-22-03	2	3323

For the ontologies ENVO, SWEET, FLOPO and PTO, we created the reference alignments for the tasks following the same procedure as in former editions. Reference files were produced using a hybrid approach consisting of (1) a 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 to this track in 2018 [4], namely: AML, Lily, the LogMap family, POMAP and XMAP.

[13] `agclass.nal.usda.gov`

[14] `www.eionet.europa.eu/gemet`

[15] Note that SKOS thesauri conceptualize by means of instances of `skos:Concept` and not `owl:Class`. Still, the *biodiv* track is different from instance matching tracks, as in both cases concepts or classes are used to define the structure (or schema) of a semantic resource.

For the thesauri AGROVOC, NALT, GEMET and ANEETHES, we created the reference alignments using the Ontology Mapping Harvesting Tool (OMHT).[16] OMHT was developed as a standalone Java program that works with one semantic resource file pulled out from AgroPortal or BioPortal[17]. OMHT automatically extracts all declared mappings by developers inside an ontology or a thesauri source files. We used for the reference alignments only the mappings with a `skos:exactMatch` property.

The evaluation was carried out on a Windows 10 (64-bit) desktop with an Intel Core i7-4770 CPU @ 3.40GHz x 4 with 16 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 [70].

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. The track also includes an analysis of False Positives.

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[18]. It comprises 2 test cases that involve 4 biomedical ontologies cov-

[16] https://github.com/agroportal/ontology_mapping_harvester
[17] https://bioportal.bioontology.org
[18] http://www.pistoiaalliance.org/projects/ontologies-mapping/

ering the disease and phenotype domains: Human Phenotype Ontology (HP) versus 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 [31]. Table 3 summarizes the versions of the ontologies used in OAEI 2020.

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 the OAEI campaigns 2016-2020 (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 contains 2,504 correspondences, whereas the DOID-ORDO one contains 3,909 correspondences.

Systems were evaluated using the standard parameters as well as the (approximate) number of unsatisfiable classes computed using the OWL 2 EL reasoner ELK [47]. 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 [8] as detailed in [42], 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 [52], LogMap [41], or AML [60]. 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 [54], 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 [47].

3.6 Multifarm

The multifarm track [53] 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[19] and Spaten [17].

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 [62], 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 [65]. 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: *Equals, 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[20], 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,

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

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

blog post or programme). The datasets were generated and transformed using SPIMBENCH [62] 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[21] 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 Geolink Cruise

The **Geolink Cruise** track consists of matching instances from different ontologies describing the same cruise in the real-world. The datasets are collected from the Geolink project,[22] which was funded under the U.S. National Science Foundation's EarthCube initiative. The datasets and alignments are guaranteed to contain real-world use cases to solve the instance matching problem in practice. In the GeoLink Cruise dataset, there are two ontologies which are GeoLink Base Ontology (gbo) and GeoLink Modular Ontology (gmo). The data providers from different organizations populate their own data into these two ontologies. In this track, we utilize instances from two different data providers, Biological and Chemical Oceanography Data Management Office (bco-

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

[22] `https://www.geolink.org/`

Table 5. The Statistics of the Ontologies in the Geolink Cruise.

Ontology	#Class	#Object Property	#Data Property	#Individual	#Triple
gbo_bco-dmo	40	149	49	1061	13055
gbo_r2r	40	149	49	5320	27992
gmo_bco-dmo	79	79	37	1052	16303
gmo_r2r	79	79	37	2025	24798

dmo)[23] and Rolling Deck to Repository (r2r)[24] and populate all the triples related to Cruise into two ontologies. There are 491 Cruise pairs between these two datasets that are labelled by domain experts as equivalent. Some statistic information of the ontologies are listed in the Table 5. More details of this benchmark can be found in the paper [6].

3.10 Knowledge Graph

The Knowledge Graph track was run for the third 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[25] in the course of the DBkWik project [34, 33]. They cover different topics (movies, games, comics and books) and three Knowledge Graph clusters sharing the same domain e.g. star trek, as shown in Table 6.

Table 6. 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

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), 4) links to disambiguation pages were manually checked and corrected. 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_265,

[23] https://www.bco-dmo.org/
[24] https://www.rvdata.us/
[25] https://www.wikia.com/

using the SEALS client (version 7.0.5). The -o option in SEALS is used to provide the two knowledge graphs which should be matched. This decreases runtime because the matching system can load the input from local files rather than downloading it from HTTP URLs. 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.[26]

Additionally we run the matchers on three hidden test cases where the source wikis are: Marvel Cinematic Universe, Memory Alpha, and Star Wars Wiki. The target wiki is for all test cases the same. It is the lyrics wiki with 1,062,920 instances, 270 properties and 67 classes. The goal is to explore how the matchers behave on matching mostly *unrelated* knowledge graphs.

As a baseline, we employed two simple string matching approaches. The source code for these matchers is publicly available.[27]

3.11 Interactive Matching

The interactive matching track aims to assess the performance of semi-automated matching systems by simulating user interaction [56, 19, 50]. 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 [39, 19].

The interactive matching track is based on the datasets from the Anatomy and Conference tracks, which have been previously described. It relies on the SEALS client's *Oracle* class to simulate user interactions. An interactive matching system can present a collection of correspondences simultaneously to the oracle, which will tell the system whether that correspondence is correct or not. If a system presents up to three correspondences together and each correspondence presented has a mapped entity (i.e., class or property) in common with at least one other correspondence presented, the oracle counts this as a single interaction, under the rationale that this corresponds to a scenario where a user is asked to choose between conflicting candidate correspondences. To simulate the possibility of user errors, the oracle can be set to reply with a given error probability (randomly, from a uniform distribution). We evaluated systems with four different error rates: 0.0 (perfect user), 0.1, 0.2, and 0.3.

[26] http://oaei.ontologymatching.org/2020/results/knowledgegraph/matching-eval-trackspecific.zip

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

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. For systems requiring more RAM, the evaluation was carried out on a Windows 10 (64-bit) desktop with an Intel Core i7-6700 CPU @ 3.40GHz x 8 with 16GB RAM allocated. Each system was run ten times and the final result of a system for each error rate represents the average of these runs. For the Conference dataset with the ra1 alignment, precision and recall correspond to the micro-average over all ontology pairs, whereas the number of interactions is the total number of interactions for all the pairs.

3.12 Complex Matching

The complex matching track is meant to evaluate the matchers based on their ability to generate complex alignments. A complex alignment is composed of complex correspondences typically involving more than two ontology entities, such as o_1*:AcceptedPaper* $\equiv$ o_2*:Paper* $\sqcap$ o_2*:hasDecision.*o_2*:Acceptance*. In addition to last year's datasets [69], two new datasets have been added: Populated Geolink and Populated Enslaved.

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 [67]. The queries are automatically rewritten using 2 systems: that from [68] 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) [12]. 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 [21].

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. 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 [72]. 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 Catalina version 10.15.6.

The **Populated GeoLink** dataset is designed to allow alignment systems that rely on the instance data to participate over the Geolink benchmark. The instance data are from real-worlds and collected from seven data repositories in the Geolink project. More detailed information on this benchmark can be found in [73]. 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 Catalina version 10.15.6.

The **Populated Enslaved** dataset was derived from the ongoing project entitled "Enslaved: People of the Historical Slave Trade[28] and funded by The Andrew W. Mellon Foundation where the focus is on tracking the movements and details of peoples in the historical slave trade. It is composed of the Enslaved ontology and the Enslaved Wikibase repository along with the populated instance data. To the best of our knowledge, it is the first attempt to align a modular ontology to the Wikibase repository. More detailed information on this benchmark can be found in [71]. 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 Catalina version 10.15.6.

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

[28] `https://enslaved.org/`

19 participating systems. Table 7 lists the participants and the tracks in which they competed. Some matching systems participated with different variants (AML, LogMap) whereas others were evaluated with different configurations, as requested by developers (see test case sections for details).

Table 7. Participants and the status of their submissions.

System	ALIN	ALOD2Vec	AML	AMLC	AROA	ATBox	DESKMatcher	CANARD	FTRLIM	Lily	LogMap	LogMap-Bio	LogMapLt	OntoConnect	RADON	RE-miner	Silk	VeeAlign	WktMtchr	Total=19
Confidence		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	16
anatomy	●	●	●	○	○	●	●	○	○	●	●	●	●	●	○	○	○	○	●	11
conference	●	●	●	○	○	●	●	○	○	●	●	○	●	○	○	○	○	●	●	10
multifarm	○	○	●	○	○	○	○	○	○	◐	●	○	●	○	○	○	○	●	●	6
complex	○	○	○	●	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	3
interactive	●	○	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	3
largebio	○	◐	●	○	○	●	◐	○	○	○	●	●	●	○	○	○	○	○	●	8
phenotype	○	●	●	○	○	●	○	○	○	○	●	●	●	○	○	○	○	○	●	7
biodiv	○	◐	●	○	○	●	○	○	○	○	●	●	●	○	○	○	○	○	◐	7
spimbench	○	○	●	○	○	○	○	○	●	●	●	○	○	○	○	●	○	○	○	5
link discovery	○	○	●	○	○	○	○	○	○	○	○	○	○	○	●	○	●	○	○	3
geolink cruise	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	0
knowledge graph	○	●	●	○	○	●	●	○	○	○	●	●	◐	○	○	○	○	○	●	8
total	3	6	10	1	1	6	4	1	1	4	9	5	7	1	1	1	1	2	7	71

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

A number of participating systems use external sources of background knowledge, which are especially critical in matching ontologies in the biomedical domain. LogMap-Bio uses BioPortal as mediating ontology provider, that is, it retrieves from BioPortal the most suitable top-10 ontologies for each matching task. LogMap uses normalizations and spelling variants from the general (biomedical) purpose SPECIALIST Lexicon. AML has three sources of background knowledge which can be used as mediators between the input ontologies: the Uber Anatomy Ontology (Uberon), the Human Disease Ontology (DOID) and the Medical Subject Headings (MeSH). XMAP and Lily use a dictionary of synonyms (pre)extracted from the UMLS Metathesaurus. In addition Lily also uses a dictionary of synonyms (pre)extracted from BioPortal.

4.2 Anatomy

The results for the Anatomy track are shown in Table 8. Of the 11 systems participating

Table 8. 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	29	1471	0.956	0.941	0.927	0.81	√
Lily	706	1517	0.901	0.901	0.902	0.747	-
LogMapBio	1005	1544	0.885	0.893	0.902	0.74	√
LogMap	7	1397	0.918	0.88	0.846	0.593	√
Wiktionary	65	1194	0.956	0.842	0.753	0.346	-
ALIN	1182	1107	0.986	0.832	0.72	0.382	√
LogMapLite	2	1147	0.962	0.828	0.728	0.288	-
ATBox	192	1030	0.987	0.799	0.671	0.129	-
ALOD2Vec	236	1403	0.83	0.798	0.768	0.386	-
OntoConnect	248	1012	0.996	0.797	0.665	0.136	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
DESKMatcher	391	2002	0.472	0.537	0.623	0.023	-

in the Anatomy track, 10 achieved an F-measure higher than the StringEquiv baseline. Three systems were first time participants (ATBox, OntoConnect, and DESKMatcher). 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 ALIN which increased in precision (from 0.974 to 0.986), recall (from 0.698 to 0.72), recall+ (from 0.365 to 0.382), F-measure (from 0.813 to 0.832), and size (from 1086 to 1107), and Lily that increased in precision (from 0.873 to 0.901), recall (from 0.796 to 0.902), recall+ (from 0.52 to 0.747), F-measure (from 0.833 to 0.901), and size (from 1381 to 1517). In terms of run time, 4 out of 11 systems computed an alignment in less than 100 seconds, a ratio which is similar to 2019 (5 out of 12). LogMapLite remains the system with the shortest runtime. Regarding quality, AML remains the system with the highest F-measure (0.941) and recall+ (0.81), but 3 other systems obtained an F-measure above 0.88 (Lily, LogMapBio, 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.

4.3 Biodiversity and Ecology

Four systems participating this year did participate to this track last year as well: AML and the LogMap family systems (LogMap, LogMapBio and LogMapLT). Three are new participants: ATBox, ALOD2Vec and Wiktionary. The newcomer ATBox did not register explicitly to the track but could cope with at least one task so we did include its results. As in the previous edition, we used precision, recall and F-measure to evaluate the performance of the participating systems. The results for the Biodiversity and Ecology track are shown in Table 9.

In comparison to previous years, we observed a decrease in the number of systems that succeeded to generate alignments for the ENVO-SWEET and FLOPO-PTO tasks. Basically, except of AML and the LogMap variants, only ATBox could cope with the

tasks with fair results. ALOD2Vec and Wiktionary generated a similar, huge set of non meaningful mappings with a very low F-measure as shown in Table 9.

Table 9. Results for the Biodiversity & Ecology track.

System	Time (s)	Number of mappings	Number of unique mappings	Precision	Recall	F-measure
FLOPO-PTO task						
LogMap	25.30	235	0	0.817	0.787	0.802
LogMapBio	450.71	236	1	0.814	0.787	0.800
AML	53.74	510	54	0.766	0.820	0.792
LogMapLt	17.02	151	0	0.987	0.611	0.755
ATBox	24.78	148	5	0.946	0.574	0.714
Wiktionary	1935	121.632	0	0.001	0.619	0.002
ALOD2Vec	246.37	121.633	1	0.001	0.619	0.002
ENVO-SWEET task						
AML	38.83	940	229	0.810	0.927	0.865
LogMapLt	32.70	617	41	0.904	0.680	0.776
ATBox	13.63	544	45	0.871	0.577	0.694
LogMap	35.15	440	0	0.964	0.516	0.672
LogMapBio	50.25	432	1	0.961	0.505	0.662
ANAEETHES-GEMET task						
LogMapBio	1243.15	397	0	0.924	0.876	0.899
LogMap	17.30	396	0	0.924	0.874	0.898
AML	4.17	328	24	0.976	0.764	0.857
LogMapLt	10.31	151	8	0.940	0.339	0.498
AGROVOC-NALT task						
AML	139.50	17.748	17.748	0.955	0.835	0.890

The results of the participating systems have slightly increased in terms of F-measure for both first two tasks compared to last year. In terms of run time, Wiktionary, ALOD2Vec and LogMapBio took the longer time, for the latter due to the loading of mediating ontologies from BioPortal.

For the FLOPO-PTO task, LogMap and LogMapBio achieved the highest F-measure. AML generated a large number of mappings (significantly bigger than the size of the reference alignment), those alignments were mostly subsumption ones. In order to evaluate the precision in a more significant manner, we had to calculate an approximation by manually assessing a subset of around 100 mappings, that were not present in the reference alignment. LogMapLt and ATBox achieved a high precision but the lowest recall.

Regarding the ENVO-SWEET task, AML ranked first in terms of F-measure, followed by LogMapLt and ATBox. The systems with the highest precision (LogMap and LogMapBio) achieve the lowest recall. Again here, AML generated a bigger set with a high number of subsumption mappings, it still achieved the best F-Measure for the

task. It is worth nothing that due the specific structure of the SWEET ontology, a lot of the false positives come from homonyms [45].

The ANAEETHES-GEMET and AGROVOC-NALT matching tasks have been introduced to the track this year, with the particularity of being resources developed in SKOS. Only AML could handle the files in their original format. LogMap and its variants could generate mappings for ANAEETHES-GEMET, based on ontology files after being transformed automatically into OWL. For the transformation, we made use of a source code[29] that was directly derived from AML ontology parsing module, kindly provided to us by its developers. LogMap and LogMapBio achieve the best results with LogMap processing the task in a shorter time. LogMapBio took a much longer time due to downloading 10 mediating ontologies from BioPortal, still the gain is not significant in terms of performance. The AGROVOC-NALT task has been managed only by AML. All other systems failed in generating mappings on both the SKOS and OWL versions of the thesauri. AML achieves good results and a very high precicion. It generated a higher number of mappings (around 1000 more) than the curated reference alignment. We performed a manual assessment of a subset of those mappings to reevaluate the precision and F-measure.

Overall, in this third evaluation, the results obtained from participating systems for the two tasks ENVO-SWEET and FLOPO-PTO remained similar with a slight increase in terms of F-measure compared to last year. The results of the two new tracks demonstrate systems (beside AML) are not ready to handle SKOS. Sometimes automatically transforming to OWL helps to avoid the issue, sometimes not. The number of mappings in the AGROVOC-NALT track is really a challenge and AML does not loose in performance which demonstrates that besides being the more tolerant tool in terms of format, it also scales up to large size thesauri.

4.4 Conference

The conference evaluation results using the sharp reference alignment *rar2* are shown in Table 10. 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 system's position: eight matching systems outperformed both baselines (ALIN, AML, ALOD2Vec, ATBox, LogMap, LogMapLt, VeeAlign and Wiktionary); two performed worse than both baselines (DESKMatcher and Lily). Three matchers (ALIN and Lily) do 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* [63, 64], as the last year, only three tools (ALIN, AML and LogMap) have no consistency principle violation.

As the last year we performed analysis of the *False Positives*, i.e. correspondences discovered by the tools which were evaluated as incorrect. The list of the False Positives

[29] `http://oaei.ontologymatching.org/2020/biodiv/code/SKOS2OWL.zip`

Table 10. 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.
VeeAlign	0.74	0.72	0.7	0.67	0.66	9	76	83
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	0.66	0.63	0.58	0.54	0.52	7	133	27
ATBox	0.58	0.58	0.57	0.56	0.56	10	192	52
LogMapLt	0.68	0.62	0.56	0.5	0.47	5	96	25
ALIN	0.82	0.69	0.56	0.48	0.43	0	2	0
ALOD2Vec	0.64	0.6	0.56	0.51	0.49	10	427	229
edna	0.74	0.66	0.56	0.49	0.45			
StringEquiv	0.76	0.65	0.53	0.45	0.41			
Lily	0.62	0.57	0.51	0.46	0.43	5	100	43
DESKMatcher	0.1	0.12	0.16	0.27	0.47	13	895	391

is available on the conference track's web page as well as further details about this evaluation. Comparing to the previous year we added the comparison of "why was an alignment discovered" assigned by us with the explanation for the alignment provided by the system itself. This year three systems generated explanations with the mappings ALOD2Vec, DESKMatcher and Wiktionary.

The Conference evaluation results using the *uncertain reference alignments* are presented in Table 11. Out of the 10 alignment systems, three (ALIN, DESKMatcher, LogMapLt) use 1.0 as the confidence value for all matches they identify. The remaining 7 systems (ALOD2Vec, AML, ATBOX, Lily, LogMap, VeeAlign, Wiktionary) have a wide variation of confidence values.

Table 11. F-measure, precision, and recall of the different matchers when evaluated using the sharp (*ra1*), discrete uncertain and continuous uncertain metrics.

	Sharp			Discrete			Continuous		
System	Prec	F-ms	Rec	Prec	F-ms	Rec	Prec	F-ms	Rec
ALIN	0.87	0.60	0.46	0.87	0.69	0.57	0.87	0.70	0.60
ALOD2Vec	0.69	0.59	0.52	0.81	0.67	0.58	0.70	0.65	0.60
AML	0.84	0.74	0.66	0.79	0.78	0.77	0.80	0.77	0.74
ATBOX	0.68	0.60	0.53	0.65	0.64	0.64	0.65	0.65	0.66
DESKMacther	0.11	0.18	0.50	0.11	0.18	0.63	0.11	0.18	0.63
Lily	0.67	0.56	0.47	1.00	0.01	0.01	0.64	0.31	0.20
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
VeeAlign	0.78	0.73	0.69	0.69	0.72	0.76	0.69	0.73	0.76
Wiktionary	0.70	0.61	0.54	0.79	0.55	0.42	0.74	0.60	0.51

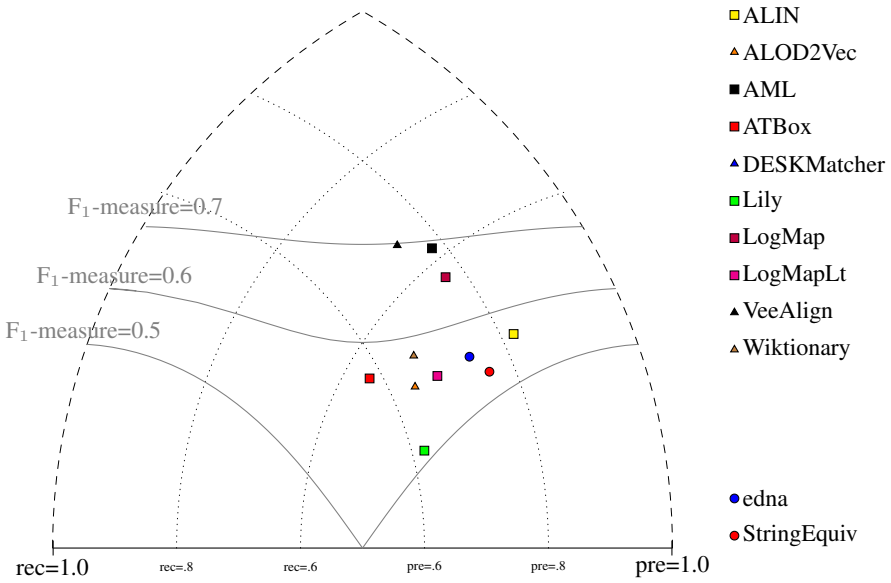

Fig. 1. Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F_1-measure are depicted by areas bordered by corresponding lines F_1-measure=0.[5|6|7].

When comparing the performance of the systems on the uncertain reference alignments versus that on the sharp version, we see that in the discrete case all systems 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 15 percent over the sharp reference alignment. This was predominantly driven by increased recall, which is a result of the presence of fewer 'controversial' matches in the uncertain version of the reference alignment.

The performance of the systems 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 (ALOD2Vec, AML, LogMap) 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 system 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 align-

ments. The exception was Lily, whose performance in discrete case decreased dramatically. ALOD2Vec, ATBOX, DESKMacther, VeeAlign are four new systems participating in this year. ALOD2Vec's performance increases 14 percent in discrete case and 11 percent in continuous case in terms of F-measure over the sharp reference alignment from 0.59 to 0.67 and 0.65 respectively, which it is mainly driven by increased recall. It is also interesting that the precision of ALOD2Vec increases 17 percent in discrete case over the sharp version. It is because ALOD2Vec assigns low confidence values to those pairs that don't have identical labels, which might help to remove some false positives in discrete case. ATBOX performs slightly better in both discrete and continuous cases compared to the sharp case in term of F-measure, which increases from 0.60 to 0.64 and 0.66 respectively. This is also mostly driven by increased recall. From the results, DESKMacther achieves low precision among three different versions of reference alignment in general because it assigns all matches with 1.0 confidence value even the labels of two entities have low string similarity. Reasonably, it achieves slightly better recall from sharp to discrete and continuous cases, while the precision and F-measure remain constant. VeeAlign's performance stays mostly constant from sharp to discrete and continuous in term of F-measure.

This year we conducted experiment of matching *cross-domain DBpedia ontology to OntoFarm ontologies*. In order to evaluate resulted alignments we prepared reference alignment of DBpedia to three OntoFarm ontologies (ekaw, sigkdd and confOf) as explained in [61]. This was not announced beforehand and systems did not specifically prepare for this. Out of 10 systems five managed to match DBpedia to OntoFarm ontologies (there were different problems dealing with parsing of the DBpedia ontology): AML, DESKMacther, LogMap, LogMapLt and Wiktionary.

We evaluated alignments from the systems and the results are in Table 12. Additionally, we added two baselines: StringEquiv as a string matcher based on string equality applied on local names of entities which were lowercased and edna as a string editing distance matcher.

Table 12. Threshold, F-measure, precision, and recall of systems when evaluated using reference alignment for DBpedia to OntoFarm ontologies

System	Thres.	Prec.	$F_{0.5}$-m.	F_1-m.	F_1-m.	Rec.
AML	0.81	0.48	0.51	0.56	0.62	0.67
edna	0.91	0.34	0.38	0.45	0.56	0.67
StringEquiv	0	0.32	0.35	0.42	0.51	0.6
Wiktionary	0.41	0.36	0.38	0.43	0.48	0.53
LogMap	0	0.37	0.39	0.41	0.45	0.47
LogMapLt	0	0.33	0.34	0.36	0.38	0.4
DESKMatcher	0	0	0	0	0	0

We can see the systems perform almost the same as two baselines except AML which dominates with 0.56 of F1-measure. Low scores of measures show that the corresponding test cases are difficult for traditional ontology matching systems since they

mainly focus on matching of domain ontologies. It is supposed to be announced as new test cases for the conference track within OAEI 2021.

4.5 Disease and Phenotype Track

In the OAEI 2020 phenotype track 7 systems were able to complete at least one of the tasks with a 6 hours timeout. Table 13 shows the evaluation results in the HP-MP and DOID-ORDO matching tasks, respectively.

Table 13. 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	32	2,128	9	0.90	0.83	0.77	≥0	≥0.0%
LogMapBio	1,355	2,198	62	0.88	0.83	0.78	≥0	≥0.0%
AML	102	2,029	358	0.91	0.82	0.74	≥0	≥0.0%
LogMapLt	7	1,370	0	1.00	0.71	0.55	≥0	≥0.0%
ATBox	16	759	10	0.98	0.46	0.30	≥0	≥0.0%
ALOD2Vec	2,384	67,943	469	0.02	0.05	0.64	≥0	≥0.0%
Wiktionary	854	67,455	4	0.02	0.04	0.63	≥0	≥0.0%
DOID-ORDO task								
LogMapBio	2,034	2,584	147	0.95	0.75	0.63	≥0	≥0.0%
AML	200	4,781	195	0.68	0.75	0.83	≥0	≥0.0%
LogMap	25	2,330	0	0.99	0.74	0.59	≥0	≥0.0%
Wiktionary	858	7,336	5	0.48	0.63	0.90	≥3,288	≥24.1%
LogMapLt	8	1,747	10	0.99	0.61	0.44	≥0	≥0.0%
ALOD2Vec	2,809	7,805	457	0.45	0.61	0.91	≥12,787	≥93.6%
ATBox	21	1,318	17	0.99	0.50	0.33	≥0	≥0.0%

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 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, LogMapBio and AML 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. Wiktionary and ALOD2Vec suggest a very large number of correspondences in the HP-MP task with respect to the

Table 14. 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	82	3,109	442	0.81	0.84	0.88	2	0.013%
LogMap	9	2,668	33	0.87	0.84	0.81	3	0.019%
LogMapBio	1,447	2,855	88	0.83	0.83	0.83	2	0.013%
LogMapLt	9	3,458	70	0.68	0.74	0.82	5,554	36.1%
Wiktionary	14,136	4,067	507	0.60	0.71	0.86	8,128	52.8%
ATBox	41	2,807	265	0.70	0.69	0.69	9,313	60.5%
Whole FMA ontology with SNOMED large fragment (Task 4)								
LogMapBio	7,046	6,470	162	0.83	0.73	0.65	0	0.0%
LogMap	624	6,540	271	0.81	0.72	0.64	0	0.0%
AML	181	8,163	2,818	0.69	0.70	0.71	0	0.0%
Wiktionary	24,379	2,034	227	0.78	0.34	0.22	989	3.0%
LogMapLt	15	1,820	26	0.85	0.33	0.21	974	2.9%
ATBox	54	1,880	124	0.80	0.33	0.21	958	2.9%
Whole NCI ontology with SNOMED large fragment (Task 6)								
AML	381	14,196	2,209	0.86	0.77	0.69	$\geq$535	$\geq$0.6%
LogMap	719	13,230	105	0.87	0.75	0.65	$\geq$1	$\geq$0.001%
LogMapBio	4,069	13,495	929	0.83	0.71	0.63	$\geq$0	$\geq$0.0%
LogMapLt	18	12,864	525	0.80	0.66	0.57	$\geq$72,865	$\geq$87.1%
Wiktionary	18,361	13,668	1,188	0.77	0.66	0.58	$\geq$68,466	$\geq$81.8%
ATBox	75	10,621	245	0.87	0.64	0.51	$\geq$65,543	$\geq$78.3%

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 and ALOD2Vec in the DOID-ORDO task.

4.6 Large Biomedical Ontologies

In the OAEI 2020 Large Biomedical Ontologies track, 8 systems were able to complete at least one of the tasks within a 6 hours timeout. Six systems were able to complete all six tasks.[30] The evaluation results for the largest matching tasks are shown in Table 14.

The top-ranked systems by F-measure were respectively: AML and LogMap in Task 2; LogMapBio and LogMap in Task 4; and AML and LogMap in Task 6. Interestingly, the use of background knowledge led to an improvement in recall from LogMapBio over LogMap in Tasks 2 and 4, 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.[31] One reason for this is that with larger ontologies there are more plausible

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

[31] `http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2020/results/`

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. 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: ALOD2Vec and DESKMatcher.

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). 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 [52], the repair module of LogMap (LogMap-Repair) [41] or the repair module of AML [60], which have worked well in practice [43, 29].

4.7 Multifarm

This year, 6 systems registered to participate in the MultiFarm track: AML, Lily, LogMap, LogMapLT, Wiktionary and VeeAlign. This number slightly increases with respect to the last campaign (5 in 2019, 6 in 2018, 8 in 2017, 7 in 2016, 5 in 2015, 3 in 2014, 7 in 2013, and 7 in 2012). Lily has generated empty alignments so there are no results to be reported.

The tools heavily rely on the lexical matching approach with the exception of VeeAlign system which adopts a deep learning approach. *VeeAlign* uses a supervised deep learning approach to discover alignments proposing a two-step model with multi-facted context representation to produce contextualised representations of concepts, which aids alignment based on semantic and structural properties of an ontology. *AML* employs lexical matching techniques using a translation module, with an emphasis on the use of background knowledge.The tool also includes structural components for both matching and filtering steps and features a logical repair algorithm. *Lily* matcher measures the literal similarity between ontologies on the extracted semantic subgraph and follows structure-based methods, background knowledge and document matching technologies. *Logmap* uses a lexical inverted index to compute the initial set of mappings which are then supported by logic based extractions with built-in reasoning and repair diagnosis capabilities. On the other hand *LogMapLt* (Logmap "lightweight") essentially only applies (efficient) string matching techniques for a lightweigh and fast computation. *Wiktionary* matcher is based on an online lexical resource, namely Wiktionary but also utilizes the schema matching and produces an explanation for the discovered correspondence. The reader can refer to the OAEI papers for a detailed description of the strategies adopted by each system.

The Multifarm evaluation results based on the blind dataset are presented in Table 15. 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). The detailed results can be investigated on the page of multifarm track results[32].

Table 15. 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) – ** tool run in a different environment so runtime is not reported; #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	170	55	8.25	.72 (.72)	.47 (.47)	.35 (.35)	33.65	.94 (.96)	.28 (.28)	.17 (.17)
LogMap	43	55	6.64	.73 (.72)	.37 (.37)	.25 (.25)	46.62	.95 (.97)	.42 (.43)	.28 (.28)
LogMapLt	17	23	1.15	.34 (.35)	.04 (.09)	.02 (.02)	95.17	.02 (.02)	.01 (.03)	.01 (.01)
VeeAlign	**	54	2.53	.73 (.77)	.15 (.15)	.09 (.09)	11.98	.91 (.93)	.14 (.14)	.08 (.08)
Wiktionary	1290	53	4.92	.77 (.80)	.32 (.33)	.21 (.21)	9.38	.94 (.96)	.12 (.13)	.07 (.07)

AML outperforms all other systems in terms of F-measure for task i) (same behaviour in the last campaigns). In terms of precision, Wiktionary is the system that generates the most precise alignments, followed by LogMap, VeeAlign and AML. With respect to the task ii) LogMap has the overall best performance. 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 2019, .37 in 2018, .36 in 2017, and .37 in 2016). The performance in terms of F-measure of Wiktionary also remains stable. In terms of runtime, the results are not really comparable with the ones in the last campaign considering the fact the SEALS repositories have been moved to another server with a different configuration.

Overall, the F-measure for blind tests remains relatively stable across campaigns. As observed in previous campaigns, systems still privilege precision over recall. Furthermore, the overall results in MultiFarm are lower than the ones obtained for the original English version of the Conference dataset.

[32] http://oaei.ontologymatching.org/2020/results/multifarm/index.html

4.8 Link Discovery

This year the Link Discovery track counted three participants in the Spatial test case: AML, Silk and RADON. Those were the exact same systems (and versions) that participated on OAEI 2019.

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 git (`https://hobbit-project.github.io/OAEI_2020.html`).

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 small dataset, but it is more clear that the systems need much more time to match instances from the TomTom dataset. RADON has still the best performance in most cases. AML has the next best performance and is able to handle some cases better than other systems (e.g. *Touches* and *Intersects*), however, it also hits the platform time limit in the case of *Disjoint*.

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

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

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

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

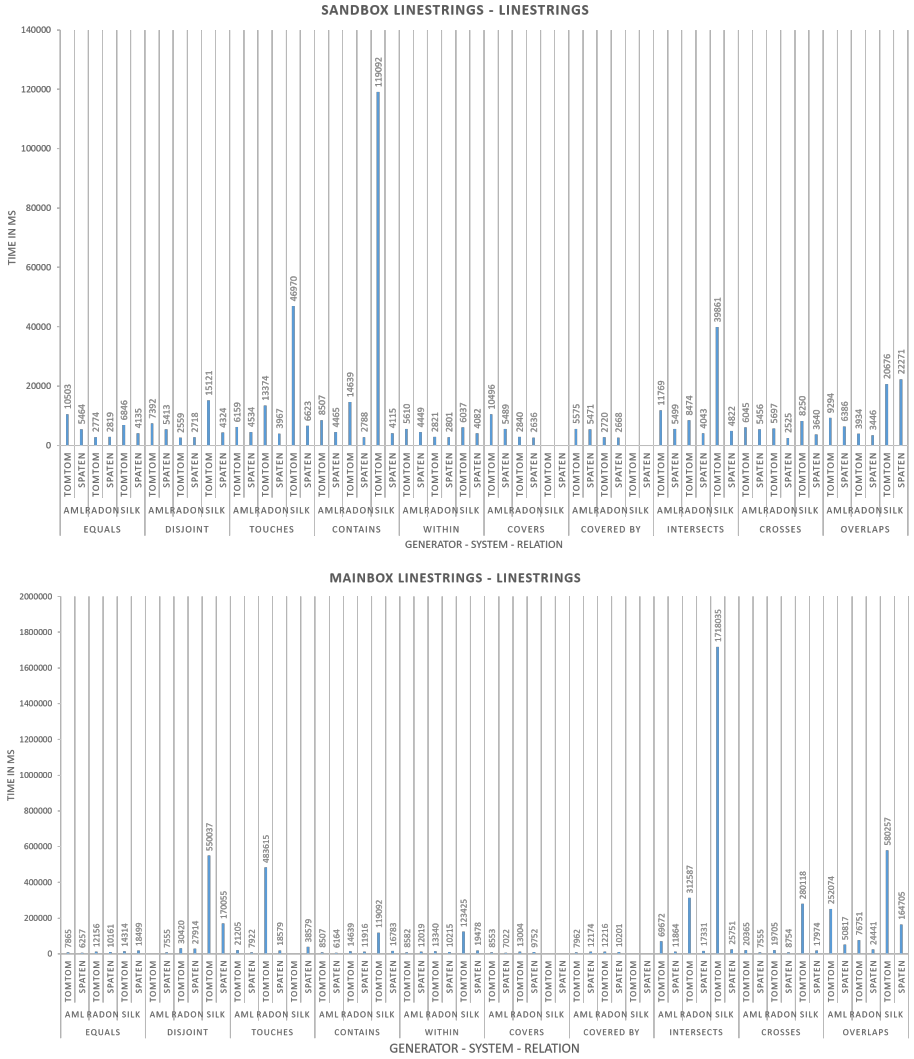

Fig. 2. Time performance for TomTom & Spaten SLL (top) and LLL (bottom) suites for AML (A), Silk (S) and RADON (R).

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.

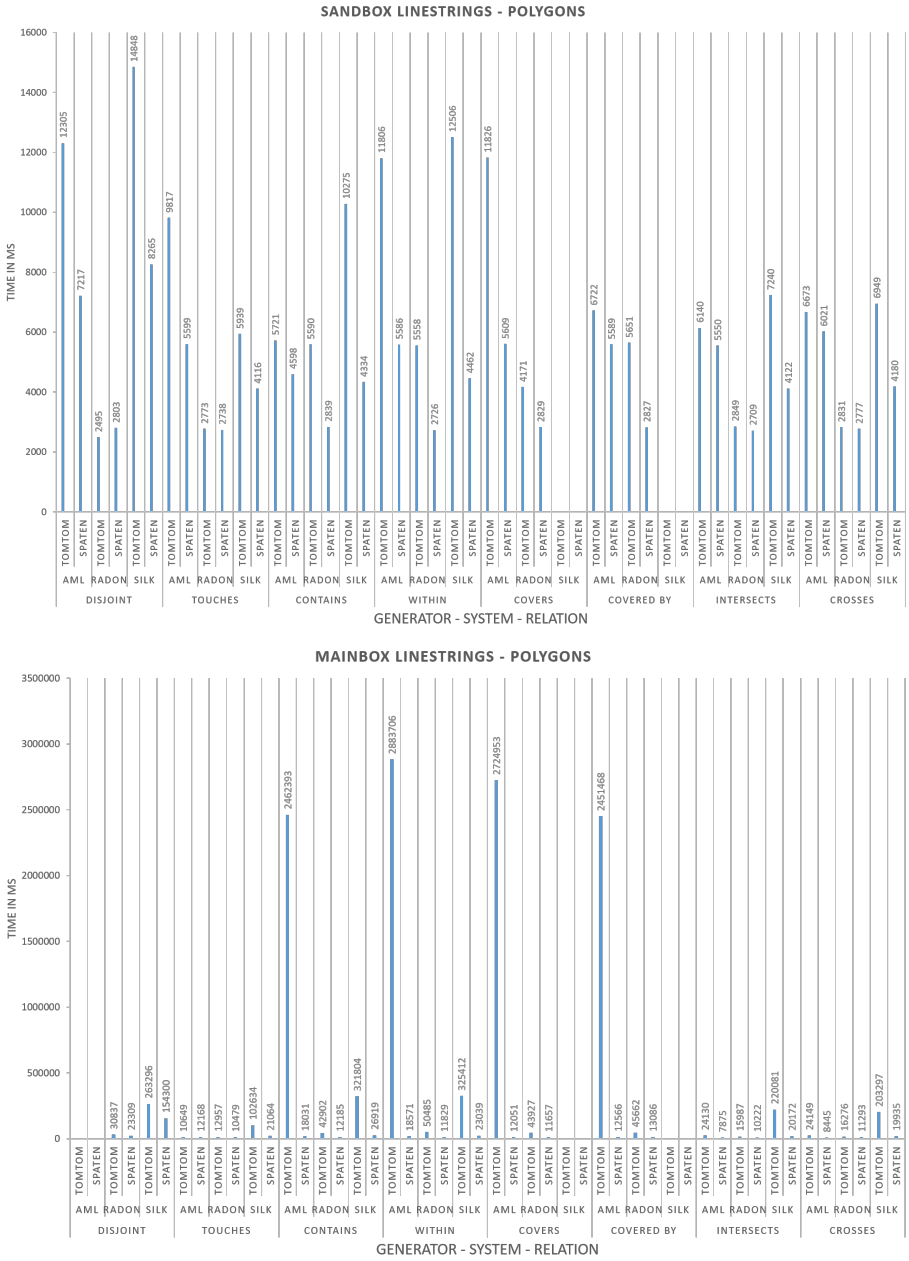

Fig. 3. Time performance for TomTom & Spaten SLP (top) and LLP (bottom) suites for AML (A), Silk (S) and RADON (R).

4.9 SPIMBENCH

This year, the SPIMBENCH track counted five participants: AML, Lily, LogMap, FTRLIM and REMiner. REMiner participated for the first time this year while AML, Lily, LogMap and FTRLIM also participated last year. The evaluation results of the track are shown in Table 16. The results can also be found in HOBBIT git (`https://hobbit-project.github.io/OAEI_2020.html`).

Table 16. Results for SPIMBENCH task.

Sandbox Dataset (380 instances, 10000 triples)				
System	**Fmeasure**	**Precision**	**Recall**	**Time (in ms)**
LogMap	0.8413	0.9382	0.7625	7483
AML	0.8645	0.8348	0.8963	6446
Lily	0.9917	0.9835	1	2050
FTRLIM	0.9214	0.8542	1	1525
REMiner	0.9983	1	0.9966	7284
Mainbox Dataset (1800 instances, 50000 triples)				
System	**Fmeasure**	**Precision**	**Recall**	**Time (in ms)**
LogMap	0.7856	0.8801	0.7094	26782
AML	0.8604	0.8385	0.8835	38772
Lily	0.9953	0.9908	1	3899
FTRLIM	0.9214	0.8558	0.9980	2247
REMiner	0.9976	0.9986	0.9966	33966

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. REMiner produces the best results (almost full) for all metrics. Lily, FTRLIM and AML had a higher recall than precision, while Lily and FTRLIM had a full recall. By contrast, REMiner and LogMap had a higher precision and lower recall, while REMiner had a full precision. AML, LogMap and REMiner had a similar run time performance.

4.10 Geolink Cruise

We evaluated all participants in the OAEI 2020. Unfortunately, none of the current alignment systems can generate the coreferences between the cruise instances in the Geolink Cruise benchmark. The state of the art alignment systems work well on finding the links with a higher string similarity or string synonyms between two objects. However, in terms of the instances with lower string similarities, or the external information is not available or very limited to help the aligning task. Another kind of algorithm is needed, like finding the relation of the instances based on the underlying structure of the graphs. We hope that system will manage this track in future years.

4.11 Knowledge Graph

We evaluated all SEALS participants in the OAEI (even those not registered for the track) on a very small matching task[33]. This revealed that not all systems were able to handle the task, and in the end, only the following systems were evaluated: ALOD2Vec, AML, ATBox, DESKMatcher, LogMapKG, LogMapLt, Wiktionary. We also evaluated LogMapBio but compared to LogMapKG it does not change the results (meaning that the external knowledge does not help in these cases which is reasonable). LogMapKG is the LogMap systems which returns TBox as well as ABox correspondences. In this year, two systems registered especially for this track but were unable to finally submit their system in time. This shows that there is a demand for this track and we plan to provide this track also next year. We hope that the system developers are able to submit the system next year. In comparison to the previous years, we have new matchers like ALOD2Vec (which produced an error in 2018), ATBox (new), and DESKMatcher (new).

What did not change over the years is that some matchers do not return a valid alignment file. The reason is the xml format of this file together with URIs in the knowledge graph containing special characters e.g. ampersand. These characters should be encoded, in order that xml parsers can process this file. Thus a post processing step is executed which tries to create a valid xml file. The resulting alignments are available for download. [34]

Table 17 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). We report the macro averaged precision, F-measure, and recall results where we do not distinguishing empty and erroneous (or not generated) alignments. The values between parentheses show the results when considering only non empty alignments.

All systems were able to generate class correspondences. In terms of F-measure, AML is still the best one and only DESKMatcher could not beat the baselines. The recall values are higher than last year (maximum of 0.77) which shows that some matchers improved and can find more class correspondences. Nevertheless there is still room for improvement and some of these class matches looks like they are not easy to find.

In the third year of this track all systems except the LogMap family are able to return property correspondences. This is a huge improvement (which happens over the years) because it makes the systems more usable in real case scenarios where a property might not be classified as owl:ObjectProperty or owl:DatatypeProperty. The systems ALOD2Vec, ATBox, and Wiktionary could achieve a F-measure of 0.95 or more which shows that property matching is easier in this track than class or instance matching.

With respect to instance correspondences, two systems (ALOD2Vec and Wiktionary) exceed the best performance of last year with an F-measure of 0.87. The margin between the baseline and the best systems is now a bit greater but still only 0.03 away. Again LogMapKG returns a much higher number of instance correspondences (29,190

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

[34] http://oaei.ontologymatching.org/2020/results/knowledgegraph/oaei2020-knowledgegraph-alignments.zip

Table 17. Knowledge Graph track results, divided into class, property, instance, and overall performance. For matchers that were not capable to complete all tasks, the numbers in parantheses denote the performance when only averaging across tasks that were completed.

System	Time (s)	# tasks	Size	Prec.	F-m.	Rec.
Class performance						
ALOD2Vec	0:13:24	5	20.0	1.00	0.80	0.67
AML	0:50:55	5	23.6	0.98	0.89	0.81
ATBox	0:16:22	5	25.6	0.97	0.87	0.79
baselineAltLabel	0:10:57	5	16.4	1.00	0.74	0.59
baselineLabel	0:10:44	5	16.4	1.00	0.74	0.59
DESKMatcher	0:13:54	5	91.4	0.76	0.71	0.66
LogMapKG	2:47:51	5	24.0	0.95	0.84	0.76
LogMapLt	0:07:19	4	23.0	0.80 (1.00)	0.56 (0.70)	0.43 (0.54)
Wiktionary	0:30:12	5	22.4	1.00	0.80	0.67
Property performance						
ALOD2Vec	0:13:24	5	76.8	0.94	0.95	0.97
AML	0:50:55	5	48.4	0.92	0.70	0.57
ATBox	0:16:22	5	78.8	0.97	0.96	0.95
baselineAltLabel	0:10:57	5	47.8	0.99	0.79	0.66
baselineLabel	0:10:44	5	47.8	0.99	0.79	0.66
DESKMatcher	0:13:54	5	0.0	0.00	0.00	0.00
LogMapKG	2:47:51	5	0.0	0.00	0.00	0.00
LogMapLt	0:07:19	4	0.0	0.00	0.00	0.00
Wiktionary	0:30:12	5	80.0	0.94	0.95	0.97
Instance performance						
ALOD2Vec	0:13:24	5	4893.8	0.91	0.87	0.83
AML	0:50:55	5	6802.8	0.90	0.85	0.80
ATBox	0:16:22	5	4858.8	0.89	0.84	0.80
baselineAltLabel	0:10:57	5	4674.8	0.89	0.84	0.80
baselineLabel	0:10:44	5	3641.8	0.95	0.81	0.71
DESKMatcher	0:13:54	5	3820.6	0.94	0.82	0.74
LogMapKG	2:47:51	5	29190.4	0.40	0.54	0.86
LogMapLt	0:07:19	4	6653.8	0.73 (0.91)	0.67 (0.84)	0.62 (0.78)
Wiktionary	0:30:12	5	4893.8	0.91	0.87	0.83
Overall performance						
ALOD2Vec	0:13:24	5	4990.6	0.91	0.87	0.83
AML	0:50:55	5	6874.8	0.90	0.85	0.80
ATBox	0:16:22	5	4963.2	0.89	0.85	0.81
baselineAltLabel	0:10:57	5	4739.0	0.89	0.84	0.80
baselineLabel	0:10:44	5	3706.0	0.95	0.81	0.71
DESKMatcher	0:13:54	5	3912.0	0.93	0.81	0.72
LogMapKG	2:47:51	5	29214.4	0.40	0.54	0.84
LogMapLt	0:07:19	4	6676.8	0.73 (0.92)	0.66 (0.83)	0.61 (0.76)
Wiktionary	0:30:12	5	4996.2	0.91	0.87	0.83

in average) than all other participants but the recall is only slighly higher (0.03 to the next best recall of 0.83).

When analyzing the confidence values of the alignments, it turns out that most matchers makes use of the range between zero and one. Only DESKMatcher, LogMapLt, and the baselines return only 1.0. Further analysis can be made by browsing to the dashboard [35] which is generated with the MELT framework [37].

Regarding runtime, LogMapKG was was the slowest system (2:47:51 for all test cases), followed by AML (0:50:55). Besides the baseline, four matchers were able to compute the alignment in under 20 minutes which is a reasonable time for this track.

In this year we also run the matchers in the hidden test cases to see how many instance correspondences they return. The systems DESKMatcher, LogMapKG, and AML (in test case starwars-lyrics) run into memory issues. Due to the fact that there is no partial nor full gold standard available for these test cases, only the number of returned instances correspondences is analyzed. In [35] we run the matchers from OAEI 2019 on these hidden test cases and manually evaluated 1,050 returned correspondences. This results in the number of matches and a approximation of the precision for each matcher and test case. Based on these values, the estimated number of true positives for each test case can be calculated. The average and maximum number of expected instance correspondences is shown in table 18 together with the number of instance correspondences returned from OAEI 2020 matchers One can see that they return 1-2 orders of magnitude more correspondences than the number of expected true positives. Especially LogMapLt returns the highest number of correspondences in the first two test cases and Wiktionary in the last test case. ATBox and AML return less correspondences and a higher precision is expected in these test cases.

Table 18. Number of instance correspondences when matching the source wiki to the lyrics wiki.

source wiki	average	max	ALOD2Vec	AML	ATBox	LogMapLt	Wiktionary
marvelcinematicuniverse	292.7	584.8	1,175	1,052	987	2,403	1,175
memoryalpha	73.6	285.5	4,546	2,106	2,817	7,195	4,547
starwars	48.5	109.1	5,697	-	3,550	2,725	5,697

4.12 Interactive matching

This year, three systems participated in the Interactive matching track. They are ALIN, AML, and LogMap. Their results are shown in Table 19 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

[35] `http://oaei.ontologymatching.org/2020/results/knowledgegraph/knowledge_graph_dashboard.html`

Table 19. 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	0.986	0.72	0.832	0.382	–	–	–	–	–	–	–
	0.0	0.988	0.856	0.917	0.623	0.988	0.856	0.917	360	953	1.0	1.0
	0.1	0.937	0.841	0.887	0.596	0.988	0.86	0.919	342	885	0.727	0.966
	0.2	0.895	0.827	0.86	0.57	0.989	0.862	0.921	337	872	0.553	0.929
	0.3	0.854	0.812	0.832	0.546	0.989	0.864	0.922	333	854	0.419	0.883
AML	NI	0.956	0.927	0.941	0.81	–	–	–	–	–	–	–
	0.0	0.972	0.933	0.952	0.822	0.972	0.933	0.952	189	189	1.0	1.0
	0.1	0.962	0.929	0.945	0.813	0.972	0.932	0.952	192	190	0.72	0.967
	0.2	0.951	0.928	0.939	0.809	0.972	0.935	0.954	212	210	0.529	0.933
	0.3	0.942	0.924	0.933	0.805	0.973	0.935	0.954	218	212	0.473	0.878
LogMap	NI	0.916	0.846	0.88	0.593	–	–	–	–	–	–	–
	0.0	0.988	0.846	0.912	0.595	0.988	0.846	0.912	388	1164	1.0	1.0
	0.1	0.967	0.831	0.894	0.567	0.971	0.803	0.879	388	1164	0.748	0.966
	0.2	0.95	0.82	0.881	0.549	0.952	0.765	0.848	388	1164	0.574	0.925
	0.3	0.938	0.818	0.874	0.543	0.927	0.723	0.812	388	1164	0.429	0.876
Conference Dataset												
ALIN	NI	0.874	0.456	0.599	–	–	–	–	–	–	–	–
	0.0	0.915	0.705	0.796	–	0.915	0.705	0.796	233	608	1.0	1.0
	0.1	0.75	0.679	0.713	–	0.928	0.736	0.821	232	597	0.581	0.988
	0.2	0.612	0.648	0.629	–	0.938	0.763	0.842	230	590	0.356	0.969
	0.3	0.516	0.617	0.562	–	0.945	0.783	0.856	227	579	0.239	0.946
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	0.843	0.682	0.754	–	0.916	0.714	0.803	242	237	0.714	0.965
	0.2	0.777	0.677	0.723	–	0.925	0.735	0.819	267	255	0.567	0.945
	0.3	0.721	0.65	0.684	–	0.929	0.742	0.825	270	253	0.452	0.879
LogMap	NI	0.818	0.59	0.686	–	–	–	–	–	–	–	–
	0.0	0.886	0.61	0.723	–	0.886	0.61	0.723	82	246	1.0	1.0
	0.1	0.851	0.6	0.703	–	0.858	0.574	0.688	82	246	0.703	0.983
	0.2	0.821	0.59	0.686	–	0.832	0.547	0.66	82	246	0.506	0.946
	0.3	0.804	0.585	0.677	–	0.817	0.522	0.637	82	246	0.385	0.909

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

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 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 [16]: 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

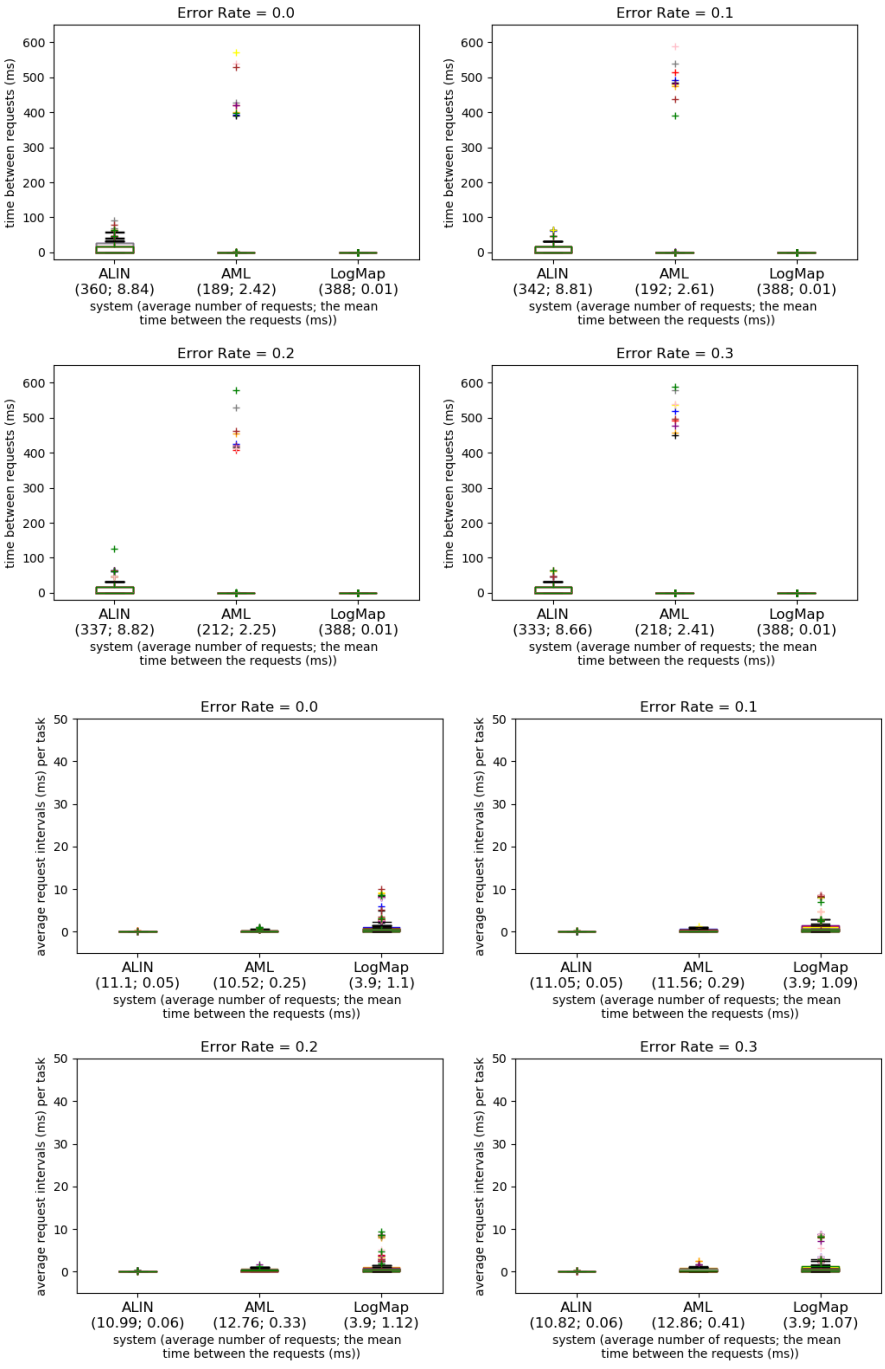

Fig. 4. Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

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 ALIN stay at a few milliseconds for most datasets. It could be the case, however, that a user would not be able to take advantage of these low response times because the task complexity may result in higher user response time (i.e., the time the user needs to respond to the system after the system is ready).

4.13 Complex Matching

Table 20. Results of the Complex Track in OAEI 2020. Populated datasets (*Pop.*) using the metrics: precision (*Prec.*), coverage (*Cov.*), relaxed precision (*R_P*), relaxed recall (*R_R*) and relaxed f-measure (*R_F*).

Matcher	Pop. Conference		Hydrography			GeoLink			Pop. GeoLink			Pop. Enslaved			Taxon	
	Prec.	Cov.	R_P	R_F	R_R	R_P	R_F	R_R	R_P	R_F	R_R	R_P	R_F	R_R	Prec.	Cov.
ALIN	**.68-.98**	.20-.28	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ALOD2Vec	.39-.78	.24-.33	-	-	-	-	-	-	-	-	-	-	-	-	.79-.96	.08-.14
AML	.59-.93	.31-.37	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AMLC	.23-.51	.26-.31	.45	.10	.05	.50	.23	.23	.50	.32	.23	.73	.40	.28	.19-.40	0
AROA	-	-	-	-	-	-	-	-	.87	**.60**	**.46**	**.80**	**.51**	**.38**	-	-
ATBox	.39-.81	.27-.36	-	-	-	-	-	-	-	-	-	-	-	-	.56-.71	.06-.11
CANARD	.25-.88	**.40-.50**	-	-	-	-	-	-	**.89**	.54	.39	.42	.19	.13	.16-.57	**.17-.36**
LogMap	.56-.96	.26-.33	.67	.10	.05	**.85**	.29	.18	.85	.29	.18	-	-	-	.54-.77	.08-.14
LogMapBio	-	-	**.70**	.10	.05	-	-	-	-	-	-	-	-	-	.50-.73	.06-.08
LogMapKG	.56-.96	.26-.33	.67	.10	.05	**.85**	.29	.18	.85	.29	.18	-	-	-	.54-.77	.08-.11
LogMapLt	.50-.87	.23-.31	.66	.10	**.06**	.69	**.36**	**.25**	.69	.36	.25	-	-	-	.25-.35	.08-.11
Wiktionary	.49-.88	.26-.35	-	-	-	-	-	-	-	-	-	-	-	-	**.89-.96**	.08-.11

Three systems were able to generate complex correspondences: AMLC, AROA, 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 20.

With respect to the Hydrography test cases, 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 and populated 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 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 CANARD 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 populated Enslaved test case, only AMLC, AROA, and CANARD can produce complex alignments. The relaxed precision of AMLC and AROA look relatively fair, while CANARD reports a lower relaxed precision. AROA found the largest number of the complex correspondences among three systems, while the AMLC outputs the largest number of the simple correspondences.

With respect to the Conference test cases the track has the same participant, AMLC, as the last year. Based on the evaluation the alignments from AMLC now conforms to the EDOAL syntax but otherwise the content of the alignment is the same.

In the Populated Conference test case, AMLC's results precision and coverage scores are lower than last year, probably because it did not take a simple reference alignment as input. CANARD's results are close to last year's. ALIN obtains the best precision score.

In the Taxon dataset, CANARD obtains the best coverage score but its precision has decreased significantly. This year, AMLC could be evaluated on this dataset ; however, the output correspondences did not cover the evaluation queries. The simple matcher obtains approximatively the same coverage score.

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

In 2020, 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. In future editions we should facilitate the participation of non-Java systems (the use of the MELT framework [36] was a step forward this year) and Machine Learning based system by providing partial alignment sets for supervised learning. Furthermore, new systems might use deep learning technology which requires specific hardware like GPUs and the like. An option would be a simple HTTP interface to allow the deployment and evaluation on different machines. The MELT framework can be easily extended with such an interface while at the same time compatibility with SEALS and HOBBIT can be retained.

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., [40]).

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). This year this has been extended by the inspection of the explanation of the correspondences provided by the systems. 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. It is, however, clear that this is not an easy task. When naively computing coherent alignments correct correspondences may be removed and incorrect ones are kept, and therefore a domain expert should be involved in the validation of different logical solutions [57, 49]. Finally, this year it was shown that matching domain ontology to cross-domain ontology is difficult task for general matching systems. While this has been done as an experiment without announcing beforehand, we suppose to announce this as new test cases within the track for next year.

With respect to the cross-lingual version of Conference, the **MultiFarm** track still attracts a few number of participants implementing specific strategies to deal with ontologies having a terminological layer in different natural languages. Despite this fact, this year new participants came with alternative strategies (i.e, deep learning) with respect to the last campaigns.

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. Comparison of the task results with embedded mappings of equivalence in the MONDO disease ontology can also be investigated in future evaluation [55].

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 domain experts. Detecting such correspondences requires the use of domain-specific core knowledge that captures biodiversity concepts. In addition this year, we put the light on the quasi total incapacity of systems to handle SKOS as input format for semantic resources to align.

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.

In the **instance matching tracks** participation increased this year for SPIMBENCH as systems became more familiar with the HOBBIT platform and had more time to do the migration. Regarding Spatial benchmark, the systems didn't have newer versions and the number of participants remained the same. Thus, the benchmark and the systems were the exact same as last year. Participation might increase next year as the systems are still updating their versions and new systems are under development. 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**, more matchers are able to match rdf:Properties and are thus better suited for real matching cases. In the third year of this track we saw a small improvement in instance alignments but the margin to the baselines is still small. In this year two new systems focused on the KG track but could not submit their systems in time. We thus expect more systems in the upcoming year.

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.

As each year, fruitful discussions at the Ontology Matching point out different directions for future improvements in OAEI. In particular, in terms of new use cases, one potential new track involves matching ontologies of units of measure (OM and QUDT) [51], in order to improve the ability of a digital twin platform to harmonise, integrate and process quantity values. Another track to be included in the next campaign is about the chemical/biological laboratory domain with strong interest from pharmaceutical companies [30, 32].

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 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).

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, Huanyu Li, Mina Abd Nikooie Pour and Ying 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).

Lu Zhou and Pascal Hitzler have been supported by the National Science Foundation under Grant No. 2033521, KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs using Spatially-Explicit AI Technologies and the Andrew W. Mellon Foundation through the Enslaved project (identifiers 1708-04732 and 1902-06575).

Beyza Yaman has been supported by the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number 13/RC/2106] and Ordnance Survey Ireland.

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). In 2020, the track was also supported by the Data to Knowledge in Agronomy and Biodiversity (D2KAB – www.d2kab.org) project that received funding from the French National Research Agency (ANR-18-CE23-0017). We would like to thank FAO AIMS and US NAL as well as the GACS project for providing mappings betweem AGROVOC and NALT. We would like to thank Christian Pichot and the ANAEE France project for providing mappings betweem ANAEETHES and GEMET.

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. Alsayed Algergawy, 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, Pavel Shvaiko, Andrea Splendiani, Élodie Thiéblin, Cássia Trojahn, Jana Vatascinová, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2019. In *Proceedings of the 14th International Workshop on Ontology Matching, Auckland, New Zealand*, pages 46–85, 2019.
6. R Amini, L Zhou, and P Hitzler. Geolink cruises: A non-synthetic benchmark for coreference resolution on knowledge graphs. In *29th ACM International Conference on Information and Knowledge Management*, 2020.
7. Benhamin Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Proc. K-Cap Workshop on Integrating Ontologies*, Banff (Canada), 2005.
8. Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.
9. Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J. Mungall, and Suzanna E. Lewis. The environment ontology: contextualising biological and biomedical entities. *Biomedical Semantics*, 4(1):43, December 2013.
10. 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.
11. 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.
12. Michelle Cheatham, Dalia Varanka, Fatima Arauz, and Lu Zhou. Alignment of surface water ontologies: a comparison of manual and automated approaches. *J. Geogr. Syst.*, 22(2):267–289, 2020.
13. Jean Clobert, André Chanzy, Jean-François Le Galliard, Abad Chabbi, Lucile Greiveldinger, Thierry Caquet, Michel Loreau, Christian Mougin, Christian Pichot, Jacques Roy, et al. How to integrate experimental research approaches in ecological and environmental studies: Anaee france as an example. *Frontiers in Ecology and Evolution*, 6:43, 2018.
14. Laurel Cooper, Ramona L. Walls, Justin Elser, Maria A. Gandolfo, Dennis W. Stevenson, Barry Smith, Justin Preece, Balaji Athreya, Christopher J. Mungall, Stefan Rensing, Manuel Hiss, Daniel Lang, Ralf Reski, Tanya Z. Berardini, Donghui Li, Eva Huala, Mary Schaeffer, Naama Menda, Elizabeth Arnaud, Rosemary Shrestha, Yukiko Yamazaki, and Pankaj Jaiswal. The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses. *Plant and Cell Physiology*, 54(2):e1, December 2012.
15. 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.
16. 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.

17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, 2nd edition, 2013.
29. 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.
30. I. Harrow et al. Ontology mapping for semantically enabled applications. *Drug Discovery Today*, 2019.

31. 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.
32. Ian Harrow, Thomas Liener, and Ernesto Jiménez-Ruiz. Ontology matching for the laboratory analytics domain. In *Proceedings of the 15th International Workshop on Ontology Matching*, 2020.
33. Sven Hertling and Heiko Paulheim. Dbkwik: A consolidated knowledge graph from thousands of wikis. In *Proceedings of the International Conference on Big Knowledge*, 2018.
34. Sven Hertling and Heiko Paulheim. Dbkwik: extracting and integrating knowledge from thousands of wikis. *Knowledge and Information Systems*, 2019.
35. Sven Hertling and Heiko Paulheim. The knowledge graph track at oaei - gold standards, baselines, and the golden hammer bias. In *The Semantic Web: ESWC 2020*, pages 343–359, 2020.
36. Sven Hertling, Jan Portisch, and Heiko Paulheim. Melt - matching evaluation toolkit. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 231–245, Cham, 2019. Springer International Publishing.
37. Sven Hertling, Jan Portisch, and Heiko Paulheim. Melt - matching evaluation toolkit. In *SEMANTICS*, 2019.
38. Robert Hoehndorf, Mona Alshahrani, Georgios V Gkoutos, George Gosline, Quentin Groom, Thomas Hamann, Jens Kattge, Sylvia Mota de Oliveira, Marco Schmidt, Soraya Sierra, et al. The flora phenotype ontology (flopo): tool for integrating morphological traits and phenotypes of vascular plants. *Journal of Biomedical Semantics*, 7(1):1–11, 2016.
39. 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.
40. Ernesto Jiménez-Ruiz, Asan Agibetov, Jiaoyan Chen, Matthias Samwald, and Valerie Cross. Dividing the Ontology Alignment Task with Semantic Embeddings and Logic-Based Modules. In *24th European Conference on Artificial Intelligence (ECAI)*, pages 784–791, 2020.
41. 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.
42. 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.
43. 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.
44. 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.
45. Naouel Karam, Abderrahmane Khiat, Alsayed Algergawy, Melanie Sattler, Claus Weiland, and Marco Schmidt. Matching biodiversity and ecology ontologies: challenges and evaluation results. *Knowl. Eng. Rev.*, 35:e9, 2020.
46. 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.

47. 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.
48. 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.
49. Patrick Lambrix. Completing and debugging ontologies: state of the art and challenges. *CoRR*, abs/1908.03171, 2019.
50. 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.
51. Francisco Martín-Recuerda, Dirk Walther, Siegfried Eisinger, Graham Moore, Petter Andersen, Per-Olav Opdahl, and Lillian Hella. Revisiting ontologies of units of measure for harmonising quantity values - A use case. In *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II*, volume 12507 of *Lecture Notes in Computer Science*, pages 551–567. Springer, 2020.
52. Christian Meilicke. *Alignment Incoherence in Ontology Matching*. PhD thesis, University Mannheim, 2011.
53. 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.
54. Boris Motik, Rob Shearer, and Ian Horrocks. Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009.
55. Christopher J Mungall, Julie A McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, Erin Foster, JP Gourdine, Julius O.B. Jacobsen, Daniel Keith, Bryan Laraway, Suzanna E. Lewis, Jeremy Nguyen Xuan, Kent Shefchek, Nicole Vasilevsky, Zhou Yuan, Nicole Washington, Harry Hochheiser, Tudor Groza, Damian Smedley, Peter N. Robinson, and Melissa A Haendel. The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, 45, 2017.
56. 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.
57. Catia Pesquita, Daniel Faria, Emanuel Santos, and Francisco M. Couto. To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. In Pavel Shvaiko, Jérôme Euzenat, Kavitha Srinivas, Ming Mao, and Ernesto Jiménez-Ruiz, editors, *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013*, volume 1111 of *CEUR Workshop Proceedings*, pages 13–24. CEUR-WS.org, 2013.
58. Robert G Raskin and Michael J Pan. Knowledge representation in the semantic web for earth and environmental terminology (sweet). *Computers & geosciences*, 31(9):1119–1125, 2005.
59. Johannes Keizer Sachit Rajbhandari. The AGROVOC Concept Scheme ; A Walkthrough. *Integrative Agriculture*, 11(5):694–699, May 2012.
60. 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.

61. Martin Šatra and Ondrej Zamazal. Towards matching of domain ontologies to cross-domain ontology: Evaluation perspective. In *Proceedings of the 19th International Workshop on Ontology Matching*, 2020.
62. 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.
63. 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.
64. Alessandro Solimando, Ernesto Jimenez-Ruiz, and Giovanna Guerrini. Minimizing conservativity violations in ontology alignments: Algorithms and evaluation. *Knowledge and Information Systems*, 2016.
65. Christian Strobl. *Encyclopedia of GIS*, chapter Dimensionally Extended Nine-Intersection Model (DE-9IM), pages 240–245. Springer, 2008.
66. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the Workshop on Evaluation of Ontology-based Tools (EON), Hiroshima (JP)*, 2004.
67. Élodie Thiéblin. Do competency questions for alignment help fostering complex correspondences? In *Proceedings of the EKAW Doctoral Consortium 2018*, 2018.
68. 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.
69. 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.
70. 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.
71. L Zhou, C Shimizu, P Hitzler, A Sheill, S Estrecha, C Foley, D Tarr, and Rehberger D. The enslaved dataset: A real-world complex ontology alignment benchmark using wikibase. In *29th ACM International Conference on Information and Knowledge Management*, 2020.
72. 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.
73. Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. Geolink data set: A complex alignment benchmark from real-world ontology. *Data Intell.*, 2(3):353–378, 2020.

Linköping, Jena, Lisboa, Heraklion, Mannheim, Montpellier, Oslo, London, Berlin,
Sankt Augustin, Trento, Toulouse, Prague, Manhattan, Dublin
December 2020