Results of the Ontology Alignment Evaluation Initiative 2021*

Mina Abd Nikooie Pour[1], Alsayed Algergawy[2], Florence Amardeilh[3], Reihaneh Amini[4], Omaima Fallatah[5], Daniel Faria[6], Irini Fundulaki[7], Ian Harrow[8], Sven Hertling[9], Pascal Hitzler[4], Martin Huschka[10], Liliana Ibanescu[11], Ernesto Jiménez-Ruiz[12,13], Naouel Karam[14,15], Amir Laadhar[16], Patrick Lambrix[1,17], Huanyu Li[1], Ying Li[1], Franck Michel[18], Engy Nasr[19], Heiko Paulheim[9], Catia Pesquita[6], Jan Portisch[9], Catherine Roussey[20], Tzanina Saveta[7], Pavel Shvaiko[21], Andrea Splendiani[8], Cássia Trojahn[22], Jana Vataščinová[23], Beyza Yaman[24], Ondřej Zamazal[23], and Lu Zhou[4]

[1] Linköping University & Swedish e-Science Research Center, Linköping, Sweden
[2] Friedrich Schiller University Jena, Germany
[3] Elzeard.co, Paris, France
[4] Data Semantics (DaSe) Laboratory, Kansas State University, USA
[5] Information School, The University of Sheffield, Sheffield, UK
[6] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[7] Institute of Computer Science-FORTH, Heraklion, Greece
[8] Pistoia Alliance Inc., USA
[9] University of Mannheim, Germany
[10] Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institut, EMI, Germany
[11] AgroParisTech, UMR MIA-Paris/INRAE, France
[12] City, University of London, UK
[13] Department of Informatics, University of Oslo, Norway
[14] Fraunhofer FOKUS, Berlin, Germany
[15] Institute for Applied Informatics (InfAI), University of Leipzig, Germany
[16] Department of Computer Science, Aalborg University, Denmark
[17] University of Gävle, Sweden
[18] University Côte d'Azur, CNRS, Inria, France
[19] Freiburg Galaxy Team, University of Freiburg, Germany
[20] INRAE Centre Clermont-ARA, laboratoire TSCF, France
[21] Trentino Digitale SpA, Trento, Italy
[22] IRIT & Université Toulouse II, Toulouse, France
[23] Prague University of Economics and Business, Czech Republic
[24] ADAPT Centre, Dublin City University, Ireland

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 2021 campaign offered 13 tracks and was attended by 21 participants. This paper is an overall presentation of that campaign.

1 Introduction

The Ontology Alignment Evaluation Initiative[1] (OAEI) is a coordinated international initiative, which organizes the evaluation of an increasing number of ontology matching systems [26, 28], and which has been run for seventeen years 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 ontology 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) [63]. 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 [55, 5, 4, 1, 2, 13, 18, 15, 3, 24, 23, 22, 11, 25, 27], which this year took place virtually [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. Since last year, the MELT framework [36] has been adopted in order to facilitate the SEALS and HOBBIT wrapping and evaluation. This year, most tracks have adopted MELT as their evaluation platform.

This paper synthesizes the 2021 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.

2 Methodology

2.1 Evaluation platforms

The OAEI evaluation was carried out in one of three alternative platforms: the SEALS client, the HOBBIT platform, or the MELT framework. All of them have the goal of ensuring reproducibility and comparability of the results across matching systems. As

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

[2] http://om2021.ontologymatching.org

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

of this campaign, the use of the SEALS client and packaging format is deprecated in favor for MELT, with the sole exception of the Interactive Matching track, as simulated interactive matching is not yet supported by MELT.

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 an interface and to wrap their tools in a predefined way including all required libraries and resources.

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 [42].

The **MELT framework**[5] [36] was introduced in 2019 and is under active development. It allows the development, evaluation, and packaging of matching systems for evaluation interfaces like SEALS or HOBBIT. It further enables developers to use Python or any other programming language in their matching systems, which beforehand had been a hurdle for OAEI participants. A newly developed evaluation client[6] allows track organizers to evaluate packaged systems whereby multiple submission formats are supported such as SEALS packages or matchers implemented as Web service.

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

2.2 Submission formats

This year, three submission formats were allowed: (1) SEALS package, (2) HOBBIT, and (3) MELT Web interface. An increasing usage of other programming languages than Java and increasing hardware requirements for matching systems was identified as challenging issue in the OAEI 2020. For addressing this issue, this year, the MELT Web interface was introduced. It mainly consists of a technology-independent HTTP interface[7] which participants can implement as they wish. Alternatively, they can use the MELT framework to assist them, as it can be used to wrap any matching system as docker container implementing the HTTP interface.

This option was very popular in the 2021 campaign: 10 systems were submitted as MELT Web docker container, 5 systems were submitted as SEALS package, 3 systems were uploaded to the HOBBIT platform, and one system implemented the Web interface directly and provided hosting for the system.

2.3 OAEI campaign phases

As in previous years, the OAEI 2021 campaign was divided into three phases: preparatory, execution, and evaluation.

[4] `https://project-hobbit.eu/outcomes/hobbit-platform/`
[5] `https://github.com/dwslab/melt`
[6] `https://dwslab.github.io/melt/matcher-evaluation/client`
[7] `https://dwslab.github.io/melt/matcher-packaging/web`

In the **preparatory phase**, the test cases were provided to participants in an initial assessment period between June 15^{th} and July 31^{st}, 2021. 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 by July 31^{st} and make a preliminary evaluation by August 30^{th}. The execution phase was terminated on October 15^{th}, 2021, at which date participants had to submit the (near) final versions of their systems (SEALS-wrapped and/or HOBBIT-wrapped).

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

3 Tracks and test cases

This year's OAEI campaign consisted of 13 tracks gathering 38 test cases, all of which included OWL ontologies to align.[8] 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 and detailed in the following sections.

3.1 Anatomy

The anatomy track comprises a single test case consisting of matching two fragments of biomedical ontologies which describe the human anatomy[9] (3304 classes) and the anatomy of the mouse[10] (2744 classes). The evaluation is based on a manually curated

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

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

[10] `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	MELT/SEALS
Biodiversity & Ecology	4	=	[0 1]	open	EN	MELT
Common Knowledge Graphs	1	=	[0 1]	open	EN	MELT
Conference	1 (21)	=, <=	[0 1]	open+blind	EN	MELT/SEALS
Disease & Phenotype	2	=, <=	[0 1]	open+blind	EN	MELT
Large Biomedical ontologies	6	=	[0 1]	open	EN	MELT
Multifarm	2 (2445)	=	[0 1]	open+blind	AR, CZ, CN, DE, EN, ES, FR, IT, NL, RU, PT	MELT
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	MELT
Interactive Matching						
Interactive	2 (22)	=, <=	[0 1]	open	EN	SEALS
Complex Matching						
Complex	7	=, <=, >=	[0 1]	open+blind	EN, ES	MELT/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.

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 machine with a 5 core CPU @ 1.80 GHz with 16GB allocated RAM, using the MELT framework. For some systems, the SEALS client has been used. 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 in section 4.

3.2 Biodiversity and Ecology

The biodiversity and ecology (biodiv) track was motivated by the GFBio[11] (The German Federation for Biological Data) and AquaDiva[12] projects, which aim at providing semantically enriched data management solutions for data capture, annotation, indexing and search [44, 46]. Since OAEI 2020 edition, we partnered with the D2KAB project[13], which develops the AgroPortal[14] ontology repository, to include new matching tasks involving important thesauri (originally developed in SKOS) in agronomy and environmental sciences. The track features the three tasks also present in former editions: matching the Environment Ontology (ENVO) to the Semantic Web for Earth and Environment Technology Ontology (SWEET), the AGROVOC thesaurus to the US National Agricultural Library Thesaurus (NALT) and the General Multilingual Environmental Thesaurus (GEMET) to the Analysis and Experimentation on Ecosystems thesaurus (ANAEETHES). This year, we address the alignment of two new biological taxonomies with rather different but complementary scopes: the well-known NCBI taxonomy (NCBITAXON), and TAXREF-LD [50], a more fine-grained, manually curated taxonomy that spans French metropolitan and overseas territories. A challenging aspect is the discrepancies between (1) the size and scope of both taxonomies, and (2) the RDF model to account for taxonomy and nomenclatural information. Table 2 presents detailed information about the ontologies and thesauri used in this year OAEI edition.

Table 2. Biodiversity and Ecology track ontologies and thesauri.

Ontology/Thesaurus	Format	Version	Classes	Instances
ENVO	OWL	2021-05-19	6,566	44
SWEET	OWL	2019-10-12	4,533	-
AGROVOC	SKOS	2020-10-02	46	706,803
NALT	SKOS	2020-28-01	2	74,158
GEMET	SKOS	2020-13-02	7	5,907
ANAEETHES	SKOS	2017-22-03	2	3,323
NCBITAXON	OWL	2021-02-15	2,308,106	-
TAXREF-LD	OWL	2020-06-23 (v13.0)	266,846	-

For ENVO-SWEET, we created the reference alignment following the same procedure as in former editions. More details about the creation process can be found in [43]. For the thesauri AGROVOC, NALT, GEMET and ANEETHES, we created the reference alignments using the Ontology Mapping Harvesting Tool (OMHT).[15] OMHT automatically extracts all declared mappings by developers inside an ontology or a thesauri

[11] www.gfbio.org
[12] www.aquadiva.uni-jena.de
[13] www.d2kab.org
[14] agroportal.lirmm.fr
[15] https://github.com/agroportal/ontology_mapping_harvester

source file pulled out from AgroPortal or BioPortal[16]. For NCBITAXON and TAXREF-LD, we created the reference alignments using Silk with a configuration that computes matches based on the short scientific names (without date nor authority). The configuration (1) selects only taxa from both ontologies with a taxonomic rank that is either species or below (subspecies, varietas etc.); and (2) normalises scientific names using taxonomic domain-specific rules so as to work around most names syntactic variations. This normalisation is implemented as a Silk plugin.[17]

3.3 Common Knowledge Graphs

The new Common Knowledge Graphs track evaluates the ability of matching systems to match the schema (classes) in large cross-domain knowledge graphs such as DBpedia [8], YAGO [62] and NELL [12]. The dataset used for the evaluation is generated from DBpedia and the Never-Ending Language Learner (NELL). While DBpedia is generated from structured data in Wikipedia's articles, NELL is an automatically generated knowledge graph with entities extracted from large-scale text corpus shared on websites. The automatic extraction process is one of the aspects that make common knowledge graphs different from ontologies, as they often result in less well-formatted and cross-domain datasets.

The evaluation is based on a gold standard of class correspondences from the two knowledge graphs [29]. Those correspondences were human annotated and verified by experts. This gold standard is only a *partial gold standard*, since not every class in each knowledge graph has an equivalent class in the opposite one. To avoid over-penalising matchers that may discover reasonable matches that are not included in the partial gold standard, our evaluation ignores any predicted matches where neither of the classes in that pair exists in a true positive pair with another class in the reference alignments. With the respect to the reference alignment, matching systems were evaluated using standard precision, recall and f-measure. The evaluation was carried out on a Linux virtual machine with 128 GB of RAM and 16 vCPUs (2.4 GHz) processors. The evaluation was performed using MELT for matchers wrapped using both SEALS, and the web packaging via Docker. As baseline, we utilize a simple string matcher which is available through MELT.

3.4 Conference

The conference track feature two test cases. The main test case 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 [64]. This year we prepared a second test case consisting of a suite of three tasks of matching DBpedia ontology (filtered to the dbpedia namespace) and three ontologies from the conference domain.

For the main test case the track uses several reference alignments for evaluation: the old (and not fully complete) manually curated open reference alignment, *ra1*; an

[16] `https://bioportal.bioontology.org`
[17] `https://github.com/frmichel/taxrefmatch-silk-plugin`

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 second test case contains open reference alignment and systems were evaluated using the standard metrics.

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.

3.5 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 covering 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 [32]. Table 3 summarizes the versions of the ontologies used in OAEI 2021.

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 2016-2021 (with vote=3). Note that systems participating with

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

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 in the OAEI 2021 thus contains 2,570 correspondences, whereas the DOID-ORDO one contains 3,967 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 [45]. 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.

3.6 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 [9] as detailed in [40], 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 [48], LogMap [39], or AML [56]. 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 [51], 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 [45].

3.7 Multifarm

The multifarm track [49] 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. In 2021, alignments have been manually evaluated by domain experts. The evaluation is blind. The systems have been executed on a Ubuntu Linux machine configured with 32GB of RAM running under a Intel Core CPU 2.00GHz x8 cores.

3.8 Link Discovery

The Link Discovery track features Spatial test case this year, 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 **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 [61]. 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

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

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] and run all the systems using a single core in order to enable a fair comparison of the systems participating in this track. But we can not fail to mention that Silk and DS-JedAI have a multi core version as well as that DS-JedAI's time performance also includes Spark start-up time.

The evaluation was carried out using the HOBBIT platform.

3.9 SPIMBENCH

The **SPIMBENCH** track consists of matching instances that are found to refer to the same real-world entity corresponding to a creative work (that can be a news item, blog post or programme). The datasets were generated and transformed using SPIMBENCH [58] 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.

3.10 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

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

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

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

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-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].

3.11 Knowledge Graph

The Knowledge Graph track was run for the fourth 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 [35, 34]. 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.

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. For evaluating all possible submission formats, MELT framework is used. The corresponding code for evaluation can be found on Github[26].

[23] https://www.bco-dmo.org/
[24] https://www.rvdata.us/
[25] https://www.wikia.com/
[26] https://github.com/dwslab/melt/tree/master/examples/kgEvalCli

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

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 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.

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

3.12 Interactive Matching

The interactive matching track aims to assess the performance of semi-automated matching systems by simulating user interaction [53, 19, 47]. 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 [37, 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.

[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 computer with an AMD Ryzen 7 5700G 3.80 GHz CPU and 32GB RAM, with 10GB of max heap space allocated to java.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.

3.13 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*.

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

The **Hydrography** dataset consists of matching four different source ontologies (hydro3, hydrOntology-translated, hydrOntology-native, and cree) to a single target ontology (SWO) [14]. 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 [66]. Evaluation was done in the same way as with the Hydrography dataset.

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 real-world data and collected from seven data repositories in the Geolink project. More detailed information on this benchmark can be found in [67]. Evaluation was done in the same way as with the Hydrography dataset.

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 [65]. Evaluation was done in the same way as with the Hydrography dataset.

The **Taxon** dataset is composed of four knowledge bases containing knowledge about plant taxonomy: AgronomicTaxon, AGROVOC, TAXREF-LD and DBpedia. The alignment systems have been executed on a Ubuntu Linux machine configured with 32GB of RAM running under a Intel Core CPU 2.00GHz x8 cores. All measurements are based on a single run.

4 Results and Discussion

4.1 Participation

Following an initial period of growth, the number of OAEI participants has remained approximately constant since 2012, at slightly over 20. This year we count with 21 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). The following sections summarise the results for each track.

4.2 Anatomy

The results for the Anatomy track are shown in Table 8. Of the 15 systems participating in the Anatomy track, 13 achieved an F-measure higher than the StringEquiv baseline. Five systems were first time participants (TOM, Fine-TOM, LSMatch, OTMapOnto and AMD). 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 decreased in precision (from 0.986 to 0.983) and increased in size (from 1107 to 1119) and recall+ (from 0.382 to 0.438), and LogMapBio which decreased in precision (from 0.885 to 0.874) and increased in size (from 1544 to 1586), recall (from 0.902 to 0.914) and recall+ (from 0.74 to 0.773). In terms of run time, 6 out of 15 systems computed an alignment in less than 100 seconds. 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. Three systems produced coherent alignments.

[28] https://enslaved.org/

Table 7. Participants and the status of their submissions.

System	ALIN	ALOD2Vec	AMD	AML	AMLC	AROA	ATMatcher	DS-JedAI	Fine-TOM	GMap	KGMatcher	Lily	LogMap	LogMap-Bio	LogMapLt	LSMatch	OTMapOnto	RADON	Silk	TOM	WktMtchr	Total=21
Confidence																						
anatomy	●	●	●	●	○	○	●	○	●	●	○	●	●	●	●	●	●	○	○	●	●	15
conference	○	●	●	●	○	○	●	○	●	●	●	●	●	○	●	●	●	○	○	●	●	14
multifarm	○	●	○	●	○	○	●	○	○	○	○	○	●	○	●	○	○	○	○	○	●	6
complex	○	○	●	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	3
interactive	●	○	○	●	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	3
largebio	○	◐	○	●	○	○	◐	○	◐	○	●	○	●	●	●	◐	◐	○	○	◐	◐	12
phenotype	○	◐	○	●	○	○	◐	○	◐	○	●	○	●	●	●	●	○	○	○	●	○	12
biodiv	○	◐	○	●	○	○	◐	○	○	○	◐	○	●	◐	●	○	○	○	○	○	○	7
mse	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	0
commonKG	○	●	●	●	○	○	●	○	○	○	●	○	●	○	○	●	●	○	○	○	●	9
spimbench	○	○	○	●	○	○	○	○	○	○	○	●	●	○	○	○	○	○	○	○	○	3
link discovery	○	○	○	●	○	○	○	●	○	○	○	○	○	○	○	○	○	●	●	○	○	4
geolink cruise	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	0
knowledge graph	○	●	◐	●	○	○	●	○	●	○	●	○	●	○	○	●	◐	○	○	●	●	11
total	2	8	5	11	1	1	8	1	5	2	6	3	10	4	6	6	5	1	1	5	6	98

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	32	1471	0.956	0.941	0.927	0.81	√
Lily	430	1517	0.901	0.901	0.902	0.747	-
LogMapBio	1043	1586	0.874	0.894	0.914	0.773	√
LogMap	7	1402	0.917	0.881	0.848	0.602	√
Fine-TOM	15068	1313	0.933	0.866	0.808	0.525	-
GMap	2362	1344	0.916	0.861	0.812	0.534	-
TOM	2647	1315	0.916	0.851	0.794	0.49	-
Wiktionary	493	1194	0.956	0.843	0.753	0.347	-
ALIN	2190	1119	0.983	0.835	0.726	0.438	-
AMD	3	1167	0.96	0.835	0.739	0.316	-
LogMapLite	2	1147	0.962	0.828	0.728	0.288	-
ALOD2Vec	261	1403	0.828	0.796	0.766	0.382	-
ATMatcher	146	1037	0.978	0.794	0.669	0.133	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
LSMatch	98	940	0.997	0.763	0.618	0.012	-
OTMapOnto	16	1903	0.646	0.72	0.811	0.515	-

4.3 Biodiversity and Ecology

This year, we have seven track participating systems. AML, ATMatcher, the LogMap family systems (LogMap, LogMapBio and LogMapLT), ALOD2Vec and KGMatcher

managed to generate an output for at least one of the track tasks. As in previous editions, 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 the previous year, roughly the same number of systems succeeded to generate alignments for the track tasks. Alongside AML and the LogMap variants, ATMatcher could cope with the tasks with fair results. KGMatcher generated a very low number of mappings and ALOD2Vec, a huge set of non meaningful mappings. Both led to a very low F-measure as shown in Table 9.

Table 9. Results for the Biodiversity & Ecology track.

System	Time (s)	Number of mappings	Precision	Recall	F-measure
ENVO-SWEET task					
AML	47	986	0.745	0.895	0.813
LogMap	13	675	0.782	0.643	0.705
LogMapLt	732	576	0.829	0.568	0.684
ATMatcher	6	572	0.817	0.569	0.671
KGMatcher	32	2	1	0.002	0.005
ANAEETHES-GEMET task					
AML	21	359	0.976	0.764	0.839
ATMatcher	8	486	0.631	0.919	0.748
LogMapLt	10	184	0.840	0.458	0.593
LogMapBio	1143	1844	0.177	0.982	0.301
LogMap	1318	1844	0.177	0.982	0.301
ALOD2Vec	103	5890	0,055	0.973	0.104
KGMatcher	32	12	0.916	0.033	0.063
AGROVOC-NALT task					
AML	196	18102	0.853	0.904	0.877

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

Regarding the ENVO-SWEET task, AML ranked first in terms of F-measure, followed by LogMap, LogMapLt and ATMatcher. The systems with the highest precision (LogMapLt and ATMatcher) achieve a similar lower recall. AML generated a bigger mapping 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 [43].

The ANAEETHES-GEMET and AGROVOC-NALT matching tasks have 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 resulting from an automatic transformation of SKOS

files into OWL. For the transformation, we made use of a source code[29] that was directly derived from the AML ontology parsing module, kindly provided to us by its developers.

For ANAEETHES-GEMET, AML achieved the best results followed by ATMatcher. LogMap and LogMapBio took a much longer time due to downloading 10 mediating ontologies from BioPortal, still the gain in terms of performance was not significant. Both systems generated a big number of mappings with a very low precision.

The AGROVOC-NALT task has been managed only by AML with good results. It generated a higher number of mappings (around 2000 more) than the curated reference alignment. We performed a manual assessment of a subset of those mappings to reevaluate the precision and F-measure. All other systems failed in generating mappings on both the SKOS and OWL versions of the thesauri. This year's newly introduced task NCBITAXON-TAXREF-LD could not be managed by any of the participating systems, due to the very large size of the considered ontologies. We plan to submit targeted subsets of the ontologies for the upcoming edition of OAEI.

Overall, in this third evaluation, the results obtained from participating systems remained similar with a slight decrease in terms of F-measure compared to last year. The results of the SKOS tasks demonstrate that systems (beside AML) are not ready to handle SKOS. By transforming the files to OWL, we could run additional systems on the tasks. Still, a native handling of SKOS and the ability to cope with SKOS thesauri specificities, like SKOS-XL lexical entities, would lead to better results.

4.4 Common Knowledge Graphs

We evaluated all the participated systems that were packaged as SEALS packages or as web services using Docker (even those not registered to participate on this new track). Although a total of 17 OAEI participants were initially evaluated, not all systems were able to handle the task. While some systems finished with an empty alignment file, others were unable to finish the task within the 12 hours timeout. Therefore, here we include the results of 9 matchers that were able to finish the task within the time limit with a non-empty alignment file which are: AML, LogMap, ALOD2Vec, OTMapOnto, KGMatcher, Wiktionary, AMD, ATmatcher, and LsMatch.

The resulted alignment files from all the participating matchers are available to download on the track's result webpage[30]. All matchers were able to discover class alignments, except for AML, which has only produced instance alignments. AMD was able to finish the task and discovered some class alignments, however, those alignments were not annotated properly for the evaluation code to process them.

Table 10 shows the result for each of the participated systems. The size column indicates the total number of class correspondences discovered by each system. With regard to f-measure, KGMatcher is the best performing matcher with an f-measure of 0.94 followed by ALOD2Vec, Wiktionary and ATmatcher that have obtained an f-measure of 0.89. In terms of precision, ALOD2Vec, Wiktionary and ATmatcher have

[29] http://oaei.ontologymatching.org/2021/biodiv/code/SKOS2OWL.zip
[30] https://oaei.ontologymatching.org/2021/results/commonKG/index.html

Table 10. Results for the Common Knowledge Graphs track

Matcher	Time	size	Precision	Recall	F1 measure
AML	00:05:19	0	0.00	0.00	0.00
LogMap	00:03:19	105	0.99	0.80	0.88
ALOD2Vec	00:04:13	103	1.00	0.80	0.89
OTMapOnto	00:08:16	123	0.90	0.84	0.87
KGMatcher	01:55:35	122	0.97	0.91	0.94
Wiktionary	00:04:32	103	1.00	0.80	0.89
AMD	00:18:27	101	0.00	0.00	0.00
ATmatcher	00:03:16	102	1.00	0.80	0.89
LsMatch	00:16:45	102	0.99	0.78	0.87
Baseline	00:00:37	78	1.00	0.60	0.75

produced the most precise alignments followed by LogMap (0.99). In terms of recall, one can also observe that most systems have privileged precision over recall, except for KGMatcher (0.91) and OTMapOnto (0.84). Both systems utilize word embeddings for the matching process to discover pairs with semantic similarity. While KGMatcher uses pre-trained word embeddings to map classes based on the similarity of their instances, the latter uses pre-trained language models to represent entities before measuring the distances between the two embeddings. Both matchers were able to discover non-trivial matches such as `placeofworship = ReligiousBuilding`, `hobby = Activity`, and `bombingevent = Attack`.

With respect to runtime, KGMatcher was the slowest system, followed by AMD and LsMatch. The shortest runtime was observed with ATmatcher and LogMap with less than 4 minutes. The size of the two knowledge graphs has caused a problem for some matchers that were unable to finish the task within the allocated time.

4.5 Conference

The conference evaluation results using the sharp reference alignment *rar2* are shown in Table 11. 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: nine systems outperformed both baselines (ALOD2Vec, AML, ATMatcher, Fine-TOM, GMap, LogMap, LogMapLt, TOM and Wiktionary); two systems performed better than StringEquiv baseline (AMD, LSMatch), and three systems performed worse than both baselines (KGMatcher, Lily and OTMapOnto). Five matchers (AMD, ATMatcher, KGMatcher, Lily[31], LSMatch) 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.

[31] Lily only outputs 15 out of 21 pair alignments.

Table 11. 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.
AML	0.78	0.74	0.69	0.65	0.62	0	39	0
LogMap	0.76	0.71	0.64	0.59	0.56	5	100	43
GMap	0.61	0.61	0.61	0.61	0.61	8	138	74
ATMatcher	0.69	0.64	0.59	0.54	0.51	1	72	8
Wiktionary	0.66	0.63	0.59	0.55	0.53	8	133	31
Fine-TOM	0.64	0.61	0.58	0.55	0.53	7	141	29
TOM	0.69	0.63	0.57	0.51	0.48	8	115	29
ALOD2Vec	0.64	0.6	0.56	0.51	0.49	10	309	205
edna	0.74	0.66	0.56	0.49	0.45			
LogMapLt	0.68	0.62	0.56	0.5	0.47	0	21	0
LSMatch	0.83	0.69	0.55	0.46	0.41	3	97	18
AMD	0.81	0.68	0.54	0.45	0.41	0	2	0
StringEquiv	0.76	0.65	0.53	0.45	0.41			
KGMatcher	0.83	0.67	0.52	0.43	0.38	0	1	0
Lily	0.62	0.57	0.51	0.46	0.43	0	2	0
OTMapOnto	0.22	0.25	0.33	0.49	0.7	15	716	593

With respect to *logical coherence* [59, 60], comparing to the last year, more systems (AMD, AML, KGMatcher, LogMap and LSMatch) have no consistency principle violation.

The Conference evaluation results using the *uncertain reference alignments* are presented in Table 12. Out of the 14 alignment systems, 6 (AMD, KGMatcher, LogMapLt, LSMatcher, OTMapOnto, TOM) use 1.0 as the confidence value for all matches they identify. The remaining 8 systems (ALOD2Vec, AML, ATMatcher, Fine-TOM, GMap, Lily, LogMap, Wiktionary) have a wide variation of confidence values.

When comparing the performance of the matchers on the uncertain reference alignments versus that on the sharp version, we see that in the discrete case all matchers, except Lily and OTMapOnto, performed the same or better in terms of F-measure (Lily's F-measure dropped almost to 0, and OTMapOnto's F-measure slightly dropped from 0.35 to 0.33). Changes in F-measure of discrete cases ranged from -1 to 16 percent over the sharp reference alignment. This was predominantly driven by increased recall, which is a result of the presence of fewer 'controversial' matches in the uncertain version of the reference alignment.

The performance of the matchers with confidence values always 1.0 is very similar regardless of whether a discrete or continuous evaluation methodology is used, because many of the matches they find are the ones that the experts had high agreement about, while the ones they missed were the more controversial matches. AML produces a fairly wide range of confidence values and has the highest F-measure under both the continuous and discrete evaluation methodologies, indicating that this system's confidence evaluation does a good job of reflecting cohesion among experts on this task. Of the

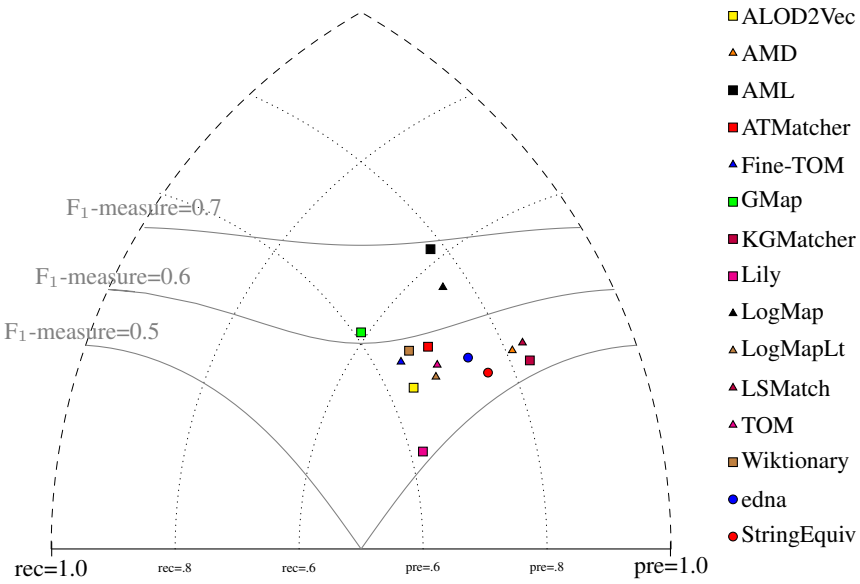

Fig. 1. Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F_1-measure are depicted by areas bordered by corresponding lines F_1-measure=0.[5|6|7].

Table 12. 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
ALOD2Vec	0.68	0.59	0.52	0.80	0.67	0.58	0.70	0.65	0.60
AMD	0.87	0.58	0.43	0.86	0.66	0.54	0.86	0.67	0.55
AML	0.84	0.74	0.66	0.79	0.78	0.77	0.80	0.77	0.74
ATMatcher	0.74	0.62	0.53	0.77	0.67	0.59	0.76	0.68	0.62
Fine-TOM	0.69	0.61	0.55	0.66	0.67	0.67	0.65	0.66	0.67
GMap	0.66	0.65	0.64	0.61	0.67	0.74	0.64	0.61	0.58
KGMatcher	0.88	0.55	0.40	0.88	0.64	0.50	0.88	0.65	0.51
Lily	0.67	0.55	0.47	1.00	0.01	0.01	0.64	0.31	0.20
LogMap	0.81	0.68	0.58	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
LSMatch	0.88	0.57	0.42	0.88	0.66	0.53	0.88	0.67	0.54
OTMapOnto	0.23	0.35	0.73	0.20	0.33	0.81	0.20	0.32	0.81
TOM	0.75	0.61	0.51	0.71	0.67	0.63	0.71	0.67	0.64
Wiktionary	0.70	0.61	0.54	0.79	0.55	0.42	0.74	0.60	0.51

remaining systems, 6 (ALOD2Vec, AML, Fine-TOM, GMap, LogMap, OTMapOnto) 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 comparing to the sharp one. This is because the matcher assigns low confidence values to some matches in which the labels are equivalent strings, which many crowdsourcers agreed with unless there was a compelling technical reason not to. This hurts recall significantly.

Overall, in comparison with last year, the F-measures of most returning matching systems essentially held constant when evaluated against the uncertain reference alignments. AMD, ATMacther, Fine-TOM, GMap, KGMatcher, LSMatcher, OTMapOnto, TOM are 8 new systems participating in this year. AMD's performance increases 14 percent in discrete case and 16 percent in continuous case in terms of F-measure over the sharp reference alignment from 0.58 to 0.66 and 0.67 respectively, which it is mainly driven by increased recall. ATMatcher, Fine-TOM, GMap, KGMatcher, and TOM perform slightly better in both discrete and continuous cases compared to sharp case in term of F-measure. This is also mostly driven by increased recall. From the results, OTMapOnto output 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, but the precision and F-measure both drop slightly.

This year we conducted experiment of matching *cross-domain DBpedia ontology* to three OntoFarm ontologies. The DBpedia ontology has been filtered to the dbpedia namespace since we merely focused on entities of DBpedia ontology (dbo). In order to evaluate resulted alignments we prepared reference alignment of DBpedia to three OntoFarm ontologies (ekaw, sigkdd and confOf) as explained in [57]. Out of 14 systems 12 (ALOD2Vec, AMD, AML, ATMatcher, Fine-TOM, KGMatcher, LogMap, LogMapLt, LSMatch, OTMapOnto, TOM and Wiktionary) managed to match dbpedia to OntoFarm ontologies.

We evaluated alignments from the systems and the results are in Table 13. 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.

Eight systems (LogMap, AML, AMD, ATMatcher, KGMatcher, LSMatch, ALOD2Vec and Wiktionary) perform better than both baselines. Four systems (TOM, Fine-TOM, LogMapLt and OTMapOnto) perform worse than both baselines. Low scores of measures show that the corresponding matching tasks are difficult for traditional ontology matching systems since they mainly focus on matching of domain ontologies.

4.6 Disease and Phenotype Track

In the OAEI 2021 phenotype track 10 systems were able to complete at least one of the tasks with a 8 hours timeout. Table 14 shows the evaluation results in the HP-MP and DOID-ORDO matching tasks, respectively.

Table 13. Threshold, F-measure, precision, and recall of systems when evaluated using reference alignment for (filtered) DBpedia to OntoFarm ontologies

System	Thres.	Prec.	$F_0.5$-m.	F_1-m.	F_2-m.	Rec.
LogMap	0.59	0.52	0.55	0.61	0.68	0.73
AML	0.81	0.5	0.53	0.59	0.67	0.73
AMD	0	0.5	0.52	0.55	0.58	0.6
ATMatcher	0.76	0.5	0.52	0.55	0.58	0.6
KGMatcher	0	0.5	0.52	0.55	0.58	0.6
LSMatch	0	0.5	0.52	0.55	0.58	0.6
ALOD2Vec	0.67	0.41	0.44	0.49	0.55	0.6
Wiktionary	0.67	0.41	0.44	0.49	0.55	0.6
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
TOM	1	0.29	0.33	0.4	0.53	0.67
Fine-TOM	1	0.26	0.29	0.36	0.48	0.6
LogMapLt	0	0.23	0.26	0.34	0.48	0.67
OTMapOnto	0	0.04	0.05	0.08	0.16	0.73

Table 14. 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	69	2,136	5	0.90	0.82	0.75	≥0	≥0.0%
LogMapBio	2,508	2,285	125	0.86	0.81	0.76	≥0	≥0.0%
AML	117	2,029	357	0.91	0.80	0.72	≥0	≥0.0%
ATMatcher	28	769	19	0.98	0.45	0.30	≥0	≥0.0%
LogMapLt	21	725	1	1.00	0.44	0.28	≥0	≥0.0%
LSMatch	2,366	685	0	1.00	0.42	0.27	≥0	≥0.0%
Fine-TOM	306	2,997	1,148	0.11	0.12	0.13	≥0	≥0.0%
TOM	306	2,493	676	0.12	0.12	0.12	≥0	≥0.0%
ALOD2Vec	3,107	67,943	66,411	0.02	0.05	0.63	≥0	≥0.0%
KGMatcher	13	3	0	1.00	0.00	0.00	≥0	≥0.0%
DOID-ORDO task								
AML	231	4,781	2,457	0.69	0.76	0.83	≥0	≥0.0%
LogMapBio	2,176	2,684	237	0.90	0.73	0.61	≥0	≥0.0%
LogMap	52	2,287	0	0.97	0.71	0.56	≥0	≥0.0%
LogMapLt	27	1,251	5	1.00	0.48	0.31	≥0	≥0.0%
LSMatch	2,749	1,193	0	1.00	0.46	0.30	≥0	≥0.0%
KGMatcher	19	338	0	1.00	0.16	0.09	≥0	≥0.0%
TOM	21	3,191	2,683	0.17	0.15	0.14	≥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 corre-

spondences 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. ALOD2Vec suggests a very large number of correspondences in the HP-MP task with respect to the other systems which suggest that it may also include many subsumption and related correspondences and not only equivalence. TOM and Fine-TOM produce a reasonable number of mappings but a very different alignment with respect to the others. KGMatcher discovers correct mappings but a very small subset. All systems produce coherent alignments using the OWL 2 EL reasoner ELK.

4.7 Large Biomedical Ontologies

In the OAEI 2021 Large Biomedical Ontologies track, 12 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.[32] The evaluation results for the largest matching tasks are shown in Table 15.

The top-ranked systems by F-measure were respectively: AML and LogMap in Task 2; LogMap and LogMapBio 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, 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.[33] One reason for this is that with larger ontologies there are more plausible correspondence candidates, and thus it is harder to attain both a high precision and a high recall. In fact, this same pattern is observed moving from the FMA-NCI to the FMA-SNOMED to the SNOMED-NCI problem, as the size of the task also increases. Another reason is that the very scale of the problem constrains the matching strategies that systems can employ: AML for example, forgoes its matching algorithms that are computationally more complex when handling very large ontologies, due to efficiency concerns. 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.

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. 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

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

[33] `http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2021/results/`

Table 15. 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	92	3,109	313	0.81	0.84	0.88	2	0.015%
LogMap	142	2,702	0	0.85	0.82	0.80	2	0.015%
LogMapBio	2,582	3,371	288	0.73	0.79	0.86	4	0.029%
LogMapLt	28	3,471	798	0.67	0.74	0.82	5,190	38.1%
KGMatcher	18	303	5	0.75	0.14	0.08	68	0.5%
Whole FMA ontology with SNOMED large fragment (Task 4)								
LogMap	761	6,463	0	0.83	0.72	0.64	0	0.0%
LogMapBio	4,921	7,377	529	0.75	0.72	0.68	0	0.0%
AML	183	8,163	2,567	0.69	0.70	0.71	0	0.0%
LogMapLt	36	1,820	31	0.85	0.33	0.21	983	3.0%
ATMatcher	77	1,890	162	0.79	0.33	0.21	962	2.9%
KGMatcher	31	252	0	0.92	0.07	0.04	0	0.0%
Whole NCI ontology with SNOMED large fragment (Task 6)								
AML	375	14,195	2,380	0.86	0.77	0.69	$\geq$0	$\geq$0.0%
LogMapBio	10,486	14,594	1,026	0.83	0.74	0.68	$\geq$0	$\geq$0.0%
LogMap	1,386	12,298	41	0.87	0.71	0.60	$\geq$0	$\geq$0.0%
LogMapLt	40	12,837	1,568	0.80	0.66	0.56	$\geq$71,454	$\geq$87.6%
KGMatcher	39	2,494	2	0.92	0.22	0.12	$\geq$19,777	$\geq$24.2%

reasoning. We encourage ontology matching system developers to develop their own repair techniques or to use state-of-the-art techniques such as Alcomo [48], the repair module of LogMap (LogMap-Repair) [39] or the repair module of AML [56], which have worked well in practice [41, 30].

4.8 Multifarm

This year, 6 systems registered to participate in the Multifarm track: ALOD2vec, AML, ATMatcher, LogMap, LogMapLT and Wiktionary. This number remains stable with respect to the last campaign (6 in 2020, 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). This year, we lost the participation of Lily and VeeAlign but we received the participation from ALOD2vec and ATMatcher replacing the missing tools.

The proposed tools are based on the lexical knowledge and dictionary approach which were applied with a mix of other approaches. Among of all the tools proposed, AML and Logmap tools provide a repair system for the links. *ALOD2vec* tool provides a neural language model approach to obtain a vector for each concept contained in the dataset which was crawled from the web of hypernymy relations as general purpose background knowledge. A system has been implemented retrieving the labels of all elements of the ontologies to be matched and linking to concepts in the background dataset to add extra context and adding to the final alignment. *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. *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 lightweight 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[34] for a detailed description of the strategies adopted by each system.

The Multifarm evaluation results based on the blind dataset are presented in Table 16 demonstrating the aggregated results for the matching tasks. They have been computed using the MELT framework without applying any threshold on the results. They are measured in terms of macro precision and recall. The results of non-specific systems are not reported here, as we could observe in the last campaigns that they can have intermediate results in tests of type ii) (same ontologies task) and poor performance in tests i) (different ontologies task).The detailed results can be investigated on the page of multifarm track results[35]. In terms of runtime, the results are not comparable to those from last year as the systems have been run in a different environment in terms of memory and number of processors. On the other hand, this year MELT framework was used instead of SEAL which was used last year.

Table 16. MultiFarm aggregated results per matcher, for each type of matching task – different ontologies. Time is measured in minutes.

	Different ontologies (i)			
System	Time(Min)	Prec.	F-m.	Rec.
ALOD2vec	10	.27	.13	.09
AML	**	.72	.47	.35
ATMatcher	113	.40	.09	.05
LogMap	9	.73	.44	.32
LogMapLt	212	.24	.04	.02
Wiktionary	157	.71	.35	.23

AML outperforms all other systems in terms of F-measure (0.47) (same behaviour in the last campaigns), followed by LogMap (0.44). In terms of precision, Logmap is the system that generates the most precise alignments, very closely followed by AML, and Wiktionary. Comparing the results from last year [55], in terms F-measure (cases of type i), AML maintains its overall performance (.47 in 2020, .45 in 2019, .46 in 2018, .46 in 2017, .45 in 2016 and .47 in 2015). On the other hand, LogMap has slightly increased its F-measure to 0.44 (.37 in 2020, .37 in 2019, .37 in 2018, .36 in 2017, and

[34] http://om2021.ontologymatching.org/

[35] http://oaei.ontologymatching.org/2021/results/multifarm/index.html

.37 in 2016). The performance in terms of f-measure of Wiktionary slightly increases F-measure to 0.35 (.32 in 2020, .31 in 2019).

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.

4.9 Link Discovery

This year the Link Discovery track counted four participants: AML, DS-JedAI, Silk and RADON. DS-JedAI participated for the first time and Silk joined with the latest version.

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 [36].

The detailed results can also be found in HOBBIT git [37]. Silk and GS-JedAI do not participate for COVERED BY and Silk also does not participate for COVERS.

In the SLL suite, RADON has the best performance in most cases except for the *Touches* and *Intersects* relations, followed by AML. DS-JedAI seems to need the most time followed by Silk.

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. On the other hand DS-JedAI, scales pretty well in larger datasets as Spark start-up time is negligible in comparison to the matching time. 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 while DS-JedAI needs the most time to complete the matchings. 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 has the second best performance. Again, DS-JedAI scales better in large datasets, thus it needs less time than Silk.

[36] In order to make the diagrams more comprehensible we have excluded the extreme values.
[37] `https://hobbit-project.github.io/OAEI_2021.html`

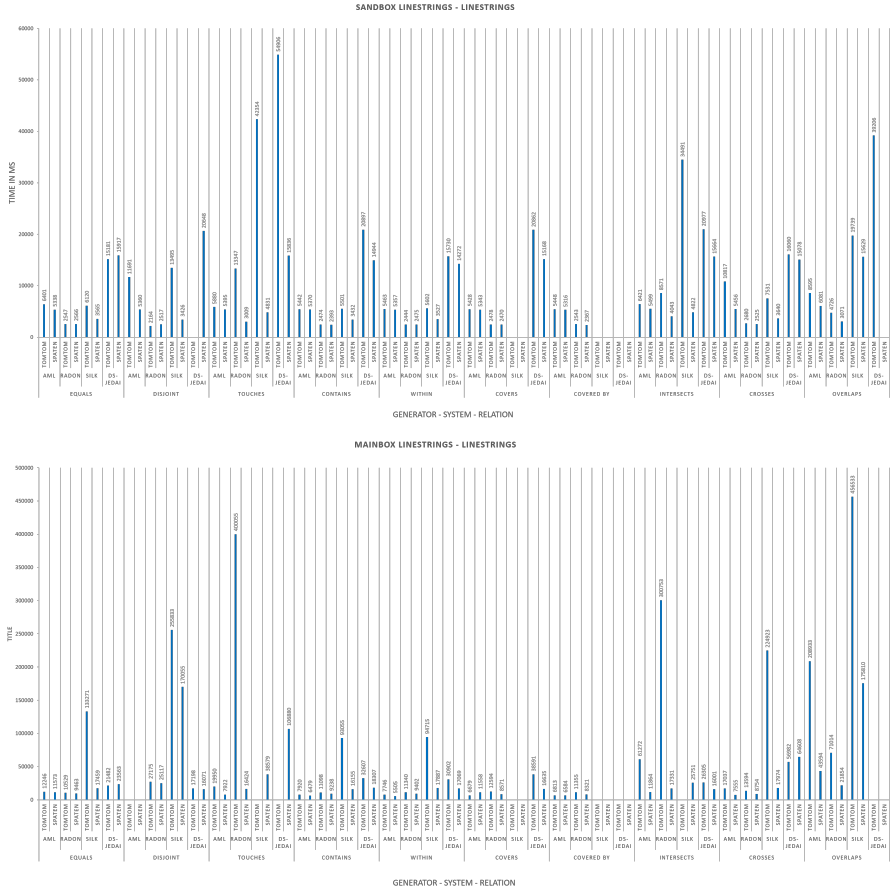

Fig. 2. Time performance for TomTom & Spaten SLL (top) and LLL (bottom) suites for AML, RADON, Silk and DS-JedAI.

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 that did not participate in the *Covers* and *Covered By* and DS-JedAI that did not participate in *Covered By* test cases. Some systems did not manage to complete some test cases, mostly *Disjoint*.

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. DS-JedAI addressed most of the tasks and scales better in larger datasets and can be improved for *Overlaps*, *Touches* and *Within*. Silk can be

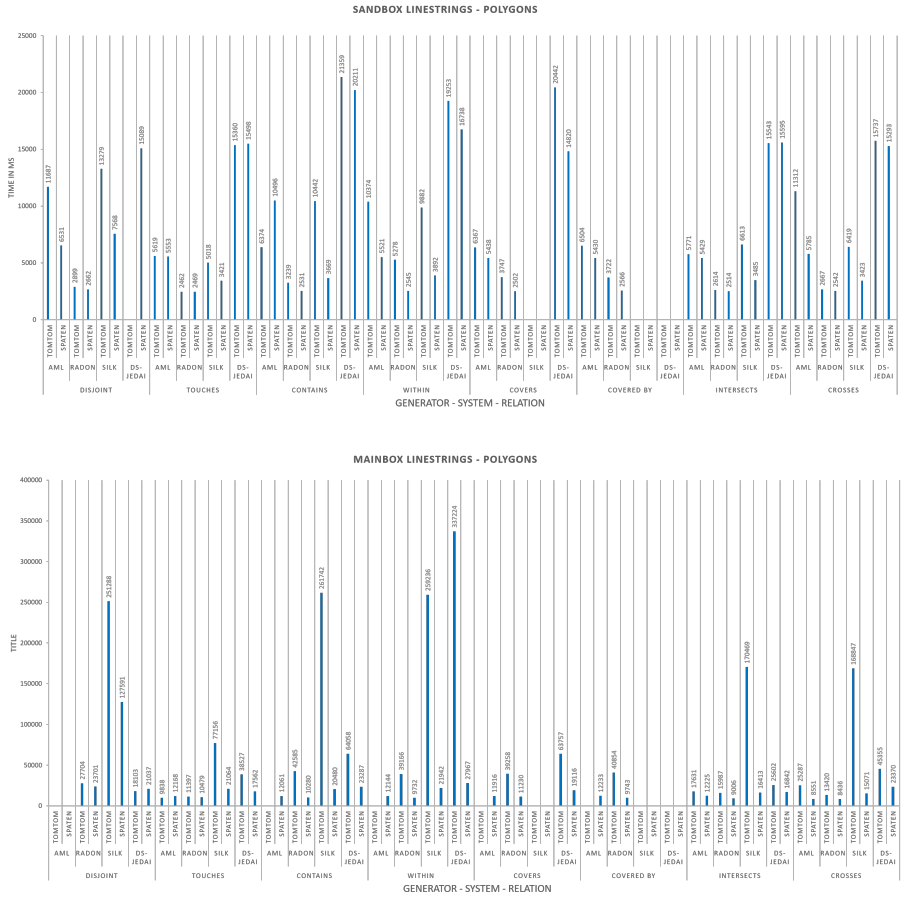

Fig. 3. Time performance for TomTom & Spaten SLP (top) and LLP (bottom) suites for AML, RADON, Silk and DS-JedAI.

improved for the *Touches*, *Intersects* and *Overlaps* relations and for the SLL and LLL tasks and for the *Disjoint* relation in SLP and LLP Tasks.

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

4.10 SPIMBENCH

This year, the SPIMBENCH track counted three participants: AML, Lily, and LogMap. All systtems participated last year. The evaluation results of the track are shown in Table 17. The results can also be found in HOBBIT git [38].

Table 17. Results for SPIMBENCH task.

Sandbox Dataset (380 instances, 10000 triples)				
System	**Fmeasure**	**Precision**	**Recall**	**Time (in ms)**
LogMap	0.8413	0.9382	0.7625	5699
AML	0.8645	0.8348	0.8963	7966
Lily	0.9917	0.9835	1	1845
Mainbox Dataset (1800 instances, 50000 triples)				
System	**Fmeasure**	**Precision**	**Recall**	**Time (in ms)**
LogMap	0.7856	0.8801	0.7094	27140
AML	0.8604	0.8385	0.8835	46517
Lily	0.9953	0.9908	1	3458

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

4.11 Geolink Cruise

We evaluated all participants in the OAEI 2021. 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.

[38] `https://hobbit-project.github.io/OAEI_2021.html`

4.12 Knowledge Graph

This year we evaluated all participants with the MELT framework to include all possible submission formats i.e. SEALS, and Web format. First, all systems are evaluated on a very small matching task[39] (even those not registered for the track). This revealed that not all systems were able to handle the task, and in the end, 11 matchers can provide results for at least one test case. This shows that over the years more and more participants adapt their systems to be able to match not only schema but also instances. When the track started in 2018, only five systems were able to finish this track, but this increased a bit to seven in 2019 and six systems in 2020 (counting the LogMap family always as one system). This year, the highest number of successful systems is reached with 11 matchers.

Similar to the previous years, some systems (AMD and AML) need a post-processing step of the resulting alignment file to be able to parse it. The reason is that the KGs in the knowledge graph track contains special characters, e.g. ampersand. These characters need to be encoded in order to parse these XML formatted files correctly. The resulting alignments are available for download [40].

Table 18. Overall performance of the systems participating in the Knowledge Graph track. This includes all types of results like class, property, and instance correspondences. 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.
overall performance						
ALOD2Vec	00:21:52	5	4990.2	0.91 (0.91)	0.87 (0.87)	0.83 (0.83)
AMD	00:37:47	2	23.0	0.40 (1.00)	0.00 (0.00)	0.00 (0.00)
AML	00:50:26	5	6874.8	0.90 (0.90)	0.85 (0.85)	0.80 (0.80)
ATMatcher	00:19:34	5	4963.4	0.89 (0.89)	0.85 (0.85)	0.81 (0.81)
BaselineAltLabel	00:11:37	5	4739.0	0.89 (0.89)	0.84 (0.84)	0.80 (0.80)
BaselineLabel	00:11:27	5	3706.0	0.95 (0.95)	0.81 (0.81)	0.71 (0.71)
Fine-TOM	14:55:09	5	4164.2	0.92 (0.92)	0.83 (0.83)	0.75 (0.75)
KGMatcher	04:55:32	5	3812.8	0.94 (0.94)	0.82 (0.82)	0.72 (0.72)
LogMap	01:04:45	5	4031.8	0.90 (0.90)	0.77 (0.77)	0.68 (0.68)
LSMatch	02:02:55	5	18.4	1.00 (1.00)	0.01 (0.01)	0.00 (0.00)
OTMapOnto	00:48:25	4	122.5	0.59 (0.73)	0.01 (0.01)	0.00 (0.01)
TOM	23:30:25	5	330.8	0.92 (0.92)	0.12 (0.12)	0.06 (0.06)
Wiktionary	00:43:18	5	4996.2	0.91 (0.91)	0.87 (0.87)	0.83 (0.83)

Table 18 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 (size). We report the macro averaged precision,

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

[40] http://oaei.ontologymatching.org/2021/results/knowledgegraph/oaei2021-knowledgegraph-alignments.zip

Table 19. Knowledge Graph track results, divided into class, property, and instance 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	00:21:52	5	20.0	1.00 (1.00)	0.80 (0.80)	0.67 (0.67)
AMD	00:37:47	2	23.0	0.40 (1.00)	0.25 (0.62)	0.18 (0.45)
AML	00:50:26	5	23.6	0.98 (0.98)	0.89 (0.89)	0.81 (0.81)
ATMatcher	00:19:34	5	25.6	0.97 (0.97)	0.87 (0.87)	0.79 (0.79)
BaselineAltLabel	00:11:37	5	16.4	1.00 (1.00)	0.74 (0.74)	0.59 (0.59)
BaselineLabel	00:11:27	5	16.4	1.00 (1.00)	0.74 (0.74)	0.59 (0.59)
Fine-TOM	14:55:09	5	19.2	1.00 (1.00)	0.80 (0.80)	0.66 (0.66)
KGMatcher	04:55:32	5	23.2	1.00 (1.00)	0.79 (0.79)	0.66 (0.66)
LogMap	01:04:45	5	19.4	0.93 (0.93)	0.81 (0.81)	0.71 (0.71)
LSMatch	02:02:55	5	18.4	1.00 (1.00)	0.78 (0.78)	0.64 (0.64)
OTMapOnto	00:48:25	4	122.5	0.59 (0.73)	0.61 (0.77)	0.64 (0.80)
TOM	23:30:25	5	19.4	1.00 (1.00)	0.83 (0.83)	0.71 (0.71)
Wiktionary	00:43:18	5	22.0	1.00 (1.00)	0.80 (0.80)	0.67 (0.67)
property performance						
ALOD2Vec	00:21:52	5	76.8	0.94 (0.94)	0.95 (0.95)	0.97 (0.97)
AMD	00:37:47	2	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AML	00:50:26	5	48.4	0.92 (0.92)	0.70 (0.70)	0.57 (0.57)
ATMatcher	00:19:34	5	78.8	0.97 (0.97)	0.96 (0.96)	0.95 (0.95)
BaselineAltLabel	00:11:37	5	47.8	0.99 (0.99)	0.79 (0.79)	0.66 (0.66)
BaselineLabel	00:11:27	5	47.8	0.99 (0.99)	0.79 (0.79)	0.66 (0.66)
Fine-TOM	14:55:09	5	29.0	0.40 (0.40)	0.39 (0.39)	0.38 (0.38)
KGMatcher	04:55:32	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
LogMap	01:04:45	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
LSMatch	02:02:55	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
OTMapOnto	00:48:25	4	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
TOM	23:30:25	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Wiktionary	00:43:18	5	79.8	0.94 (0.94)	0.95 (0.95)	0.97 (0.97)
instance performance						
ALOD2Vec	00:21:52	5	4893.4	0.91 (0.91)	0.87 (0.87)	0.83 (0.83)
AMD	00:37:47	2	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AML	00:50:26	5	6802.8	0.90 (0.90)	0.85 (0.85)	0.80 (0.80)
ATMatcher	00:19:34	5	4859.0	0.89 (0.89)	0.85 (0.85)	0.80 (0.80)
BaselineAltLabel	00:11:37	5	4674.8	0.89 (0.89)	0.84 (0.84)	0.80 (0.80)
BaselineLabel	00:11:27	5	3641.8	0.95 (0.95)	0.81 (0.81)	0.71 (0.71)
Fine-TOM	14:55:09	5	4116.0	0.92 (0.92)	0.83 (0.83)	0.76 (0.76)
KGMatcher	04:55:32	5	3789.6	0.94 (0.94)	0.82 (0.82)	0.74 (0.74)
LogMap	01:04:45	5	4012.4	0.90 (0.90)	0.78 (0.78)	0.69 (0.69)
LSMatch	02:02:55	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
OTMapOnto	00:48:25	4	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
TOM	23:30:25	5	311.4	0.91 (0.91)	0.12 (0.12)	0.06 (0.06)
Wiktionary	00:43:18	5	4894.4	0.91 (0.91)	0.87 (0.87)	0.83 (0.83)

F-measure, and recall results, where we do not distinguishing empty and erroneous (or not generated) alignments. The values in parentheses show the results when considering only non empty alignments.

In terms of F-measure, ALOD2Vec and Wiktionary still achieve 0.87 (as in previous years) and no other system can improve on that. They also return a high amount of correspondences. AML produces the largest number of correspondences (6,875) on average, however, other systems reach a higher recall at lower absolute numbers.

Regarding runtime, TOM (23:30:25) and Fine-TOM (14:55:09) were the slowest systems. This is probably due to the fact that both are transformer based systems. The computation of these models takes time on machines without any GPU support. Besides the baselines (which need around 12 minutes for all test cases) ATMatcher (00:19:34) and ALOD2Vec (00:21:52) were the fastest systems.

In table 19, the results are further distinguished between class, property, and instance correspondences. They are also averaged over all five test cases in this track. Detailed results for each test case can be found on the OAEI results page of the track[41].

All systems are able to return class correspondences. The baseline results show that it is easy to achieve a high precision when matching only based on string comparison. In terms of F-measure, AML is again the best performing system, mainly due to the high recall of 0.81. Only ATMatcher and OTMapOnto (when counting only successful test cases) have similar recall values (0.79 and 0.80, respectively). On the other end of the spectrum, AMD is the only system which could not beat the baseline in terms of F-measure. Interestingly, it is also one of the systems which focuses on the class correspondences.

Analyzing the property correspondences reveals the same observation as in previous years. Many of the systems do not match `rdf:Property`, but only handle properties which are classified into `owl:ObjectProperty` or `owl:DatatypeProperty`. This year, six systems were not capable of producing any property matches, which is a significant increase compared to 2020. ATMatcher has the highest F-measure of 0.96, closely followed by ALOD2Vec and Wiktionary with 0.95. Overall, only five systems actually returned any property correspondences.

The highest amount of correspondences in the gold standard are available for instances (15,361). Only three systems (AMD, LSMatch, and OTMapOnto) do not return instance alignments. All others score between 0.78 and 0.87 F-measure (TOM is an exception here with only 0.12). Also for the instances, it is easier to achieve a high precision (0.94 of KGMatcher) than a high recall (best value is 0.83 by ALOD2Vec and Wiktionary). The best scores for instances matching did not improve in comparison to last year.

For further analysis of the results, we also provide an online dashboard[42] generated with MELT[54]. It allows to inspect the results on a correspondence level. Due to the large amount of these correspondences (203,935), it can take some time to load the full dashboard. Once finished, it allows to analyse the distribution of confidences. AML,

[41] `http://oaei.ontologymatching.org/2021/results/knowledgegraph/index.html`

[42] `http://oaei.ontologymatching.org/2021/results/knowledgegraph/knowledge_graph_dashboard.html`

ATMatcher, Fine-Tom, LogMap, and TOM not only use one confidence of 1.0 but have different values for correspondences. The full range between zero and one is used by ALOD2Vec and Wiktionary.

4.13 Interactive matching

This year, three systems (ALIN, AML, and LogMap) participated in the Interactive matching track. Their results are shown in Table 20 and Figure 4 for both the Anatomy and Conference datasets.

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

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

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

The matching systems that participated in this track employ different user-interaction strategies. While LogMap, and AML make use of user interactions exclusively in the post-matching steps to filter their candidate correspondences, ALIN can also add new candidate correspondences to its initial set. LogMap and AML both request feedback on only selected correspondences candidates (based on their similarity 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.

Table 20. 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	0.983	0.726	0.835	0.438	–	–	–	–	–	–	–
	0.0	0.986	0.887	0.934	0.702	0.986	0.887	0.934	404	1042	1.0	1.0
	0.1	0.933	0.866	0.899	0.669	0.987	0.887	0.934	360	910	0.667	0.868
	0.2	0.883	0.846	0.864	0.639	0.987	0.885	0.933	387	970	0.548	0.926
	0.3	0.756	0.745	0.75	0.548	0.888	0.797	0.84	380	956	0.415	0.881
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	0.961	0.931	0.946	0.819	0.972	0.935	0.953	201	199	0.717	0.974
	0.2	0.952	0.927	0.939	0.811	0.972	0.934	0.953	209	204	0.591	0.928
	0.3	0.942	0.925	0.933	0.805	0.973	0.936	0.954	214	211	0.443	0.885
LogMap	NI	0.915	0.847	0.88	0.602	–	–	–	–	–	–	–
	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	0.565	0.972	0.805	0.881	388	1164	0.753	0.967
	0.2	0.947	0.823	0.881	0.552	0.949	0.759	0.844	388	1164	0.557	0.928
	0.3	0.939	0.819	0.875	0.543	0.929	0.727	0.815	388	1164	0.439	0.88
Conference Dataset												
ALIN	NI	0.874	0.456	0.599	–	–	–	–	–	–	–	–
	0.0	0.916	0.718	0.805	–	0.916	0.718	0.805	281	718	1.0	1.0
	0.1	0.72	0.677	0.698	–	0.93	0.746	0.828	272	693	0.528	0.988
	0.2	0.58	0.646	0.611	–	0.941	0.775	0.85	240	613	0.321	0.971
	0.3	0.5	0.622	0.554	–	0.947	0.79	0.861	208	536	0.23	0.956
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	0.845	0.687	0.758	–	0.916	0.717	0.804	245	239	0.726	0.966
	0.2	0.777	0.665	0.717	–	0.923	0.729	0.815	265	254	0.542	0.929
	0.3	0.724	0.65	0.685	–	0.928	0.746	0.827	272	257	0.455	0.87
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	0.847	0.595	0.699	–	0.856	0.577	0.69	82	246	0.705	0.973
	0.2	0.819	0.589	0.685	–	0.834	0.548	0.662	82	246	0.494	0.94
	0.3	0.796	0.586	0.675	–	0.81	0.516	0.63	82	246	0.365	0.912

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

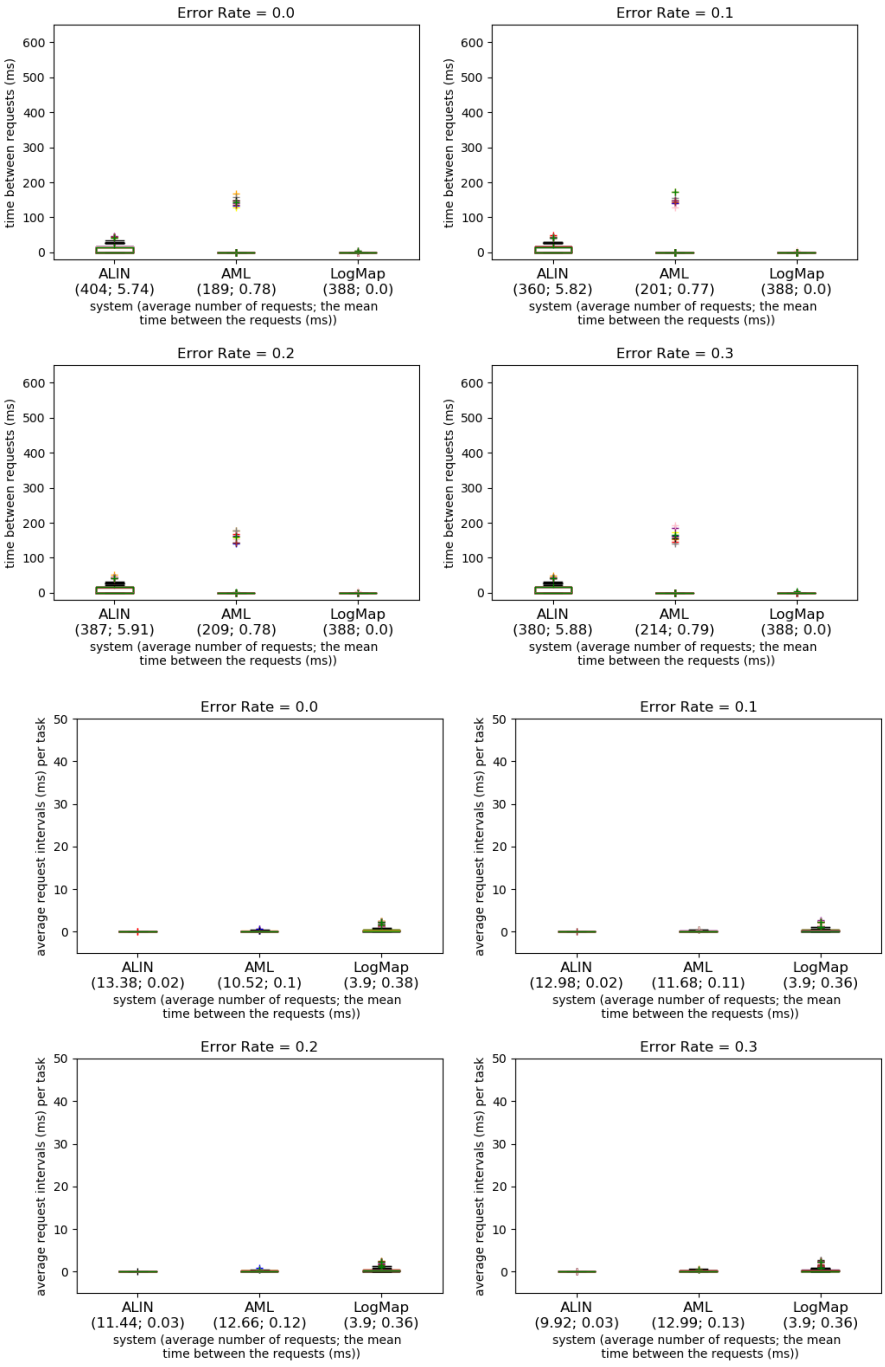

Fig. 4. Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

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 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).

4.14 Complex Matching

Three systems were able to generate complex correspondences: AMD, AMLC, and AROA. The results for the other systems are reported in terms of simple alignments. The results of the systems on four out of the five test cases are summarized in Table 21.

With respect to the Hydrography test cases, none of the systems can generate complex correspondences in this year. 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 only two alignment systems that can generate complex alignments in GeoLink Benchmark, which are AMLC and AROA. AMLC didn't find any correct complex alignment, while AROA still achieved relatively good performance. One of the reasons is that AROA is instance-based systems, which rely on the shared instances between ontologies. In other words, finding related instances between two ontologies or knowledge graphs can be helpful to improve the performance of the matching process.

In the populated Enslaved test case, besides AMLC and AROA can produce complex alignments, LogMap also can find complex correspondences this year. The relaxed

Table 21. Results of the Complex Track in OAEI 2021. 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	Hydrography			GeoLink			Pop. GeoLink			Pop. Enslaved		
	R_P	R_F	R_R	R_P	R_F	R_R	R_P	R_F	R_R	R_P	R_F	R_R
ALIN	-	-	-	-	-	-	-	-	-	-	-	-
ALOD2Vec	-	-	-	-	-	-	-	-	-	-	-	-
AMD	-	-	-	-	-	-	-	-	-	-	-	-
AML	.49	.08	.04	-	-	-	-	-	-	-	-	-
AMLC	-	-	-	.49	.30	.22	.49	.30	.22	.46	.18	.12
AROA	-	-	-	-	-	-	**.87**	**.60**	**.46**	**.80**	**.51**	**.38**
ATMatcher	-	-	-	-	-	-	-	-	-	-	-	-
Fine-TOM.	-	-	-	-	-	-	-	-	-	-	-	-
GMap	-	-	-	-	-	-	-	-	-	-	-	-
KGMatcher	-	-	-	-	-	-	-	-	-	-	-	-
Lily	-	-	-	-	-	-	-	-	-	-	-	-
LogMap	.67	.10	.05	**.85**	.29	.18	.85	.29	.18	.69	.19	.11
LogMapBio	**.70**	.10	.05	-	-	-	-	-	-	-	-	-
LogMapLt	.66	.10	**.06**	.69	**.36**	**.25**	.69	.36	.25	-	-	-
LSMatch	-	-	-	-	-	-	-	-	-	-	-	-
TOM	-	-	-	-	-	-	-	-	-	-	-	-
Wiktionary	-	-	-	-	-	-	-	-	-	-	-	-

precision of AROA and LogMap look relatively fair, while AMLC reports a lower relaxed precision than last year. AROA found the largest number of the complex correspondences among three systems, while the LogMap 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. This year AMLC delivered the alignments consisting of both, simple and complex correspondences. Within this evaluation only complex correspondences were evaluated and the results are the same as the last year.

In the Taxon dataset, only two systems applied for participating in the task, AMLC and AMD, among which only AMLC was able to produce results, with AMD not being able to parse the files. We also ran the systems generating simple alignments. Two main challenges make the alignment task difficult: i) The four taxonomic registers in the Taxon dataset adopt somewhat different approaches to model taxonomic information using instances of SKOS Concept and OWL classes. The modelling discrepancies entail that alignments should be able to "cross" modelling perspectives, e.g. aligning an OWL class with an instance of SKOS concept; ii) situations occur where all taxonomic registers are not on the same page as a result of the fact that scientific consensus about taxonomy constantly evolves. While the experts expected to evaluate the links between taxonomic entities, in many other cases, the aligned resources were entities from the vocabularies shared by several taxonomic registers (e.g. properties from SKOS or Dublin Core Terms). Although such alignments are often true, they are usually rather obvious and hence useless. Besides, although the taxonomic registers contain thousands to

hundreds of thousands of taxa each, only very little alignments were proposed between these entities. Overall, no valid complex alignments were proposed between taxa, and LogMap was the only system that seems to be able to yield simple alignments that deal with ii).

A more detailed discussion of the results of each task can be found in the OAEI page for this track[43]. 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 2021 we witnessed a healthy mix of new and returning systems. Like last year, the distribution of participants by tracks was uneven. In future editions we plan to facilitate the participation of non-Java systems (the use of the MELT framework [36] was a step forward this year) and Machine Learning based systems by providing partial alignment sets for supervised learning.

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 (see, e.g., [38]).

According to the **Conference** track there is still need for an improvement with regard to the ability of matching systems to match properties. This year we witnessed more systems (five) concerned with the logical coherence of the alignments they produce, an aspect which is critical for several semantic web applications. However, we will see next year whether there is really a growing trend. Finally, this year it was shown that matching domain ontology to cross-domain ontology is difficult task for general matching systems.

With respect to the cross-lingual version of Conference, the **MultiFarm** track still attracts too few number of participants. 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

[43] `https://oaei.ontologymatching.org/2021/complex/index.html`

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 [52].

In the **Biodiversity and Ecology track**, 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 did confirm on the one hand the inability of most systems to handle SKOS as input format and to handle very large ontologies and thesauri in the other hand. We plan to reuse techniques from the Large Biomedical Ontologies track as well as experts knowledge to provide manageable subsets.

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 16 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** tackles a challenge task and still attracts a very few number of participants. This year, domain experts have been manually evaluated the generated alignments and have been confronted with difficulties as the lack of user interfaces for manipulating complex alignments and helping understanding EDOAL. This track has also to evolve, in particular the Taxon track, considering new versions of the used resources (TaxRef-LD) and additional resources as NCBI and DBpedia.

In the **instance matching tracks** participation decreased this year for SPIMBENCH and increased for Spatial benchmark. Regarding Spatial benchmark some systems had newer versions. 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 participate in this track. Still, seven of them do not match rdf:Properties. In the fourth year of this track we saw a small improvement in instance alignments but the margin to the baselines is still small.

In the new **common knowledge graphs track**, which challenges matching systems to map the schema of large-scale, automatically constructed, and cross-domain knowledge graphs, a number of systems were able to finish the task, while others faced a problem coping with the dataset size. Some of the systems that utilize deep learning techniques such as transfer learning were not able to complete the task within the allotted time. Therefore, we expect those systems to be adapted to the task, and we look forward to having more participants in the upcoming campaign.

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 food product concepts [10]. Another track to be included in the next campaign is about the chemical/biological laboratory domain with strong interest from pharmaceutical companies [31, 33].

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 and Catia Pesquita were supported by the FCT through the LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020) and by the KATY project funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 101017453.

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).

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

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 Coun-

cil (Vetenskapsrådet, dnr 2018-04147) and the Swedish National Graduate School in Computer Science (CUGS).

Lu Zhou has 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 NFDI4BioDiversity project (Project number 442032008) and the CRC 1076 AquaDiva. In 2021, 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 between AGROVOC and NALT. We would like to thank Christian Pichot and the ANAEE France project for providing mappings between 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. Christian Bizer, Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mende, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, pages 1–5, 2012.
9. Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.
10. Patrice Buche, Cufi Julien, Stéphane Dervaux, Juliette Dibie, Liliana Ibanescu, Alrick Oudot, and Magalie Weber. A new alignment method based on foodon as pivot ontology to manage incompleteness in nutritional legacy data sources (short paper). In Janna Hastings and Frank Loebe, editors, *Proceedings of the 11th International Conference on Biomedical Ontologies (ICBO) joint with the 10th Workshop on Ontologies and Data in Life Sciences (ODLS) and part of the Bolzano Summer of Knowledge (BoSK 2020), Virtual conference hosted in Bolzano, Italy, September 17, 2020*, volume 2807 of *CEUR Workshop Proceedings*, pages 1–2. CEUR-WS.org, 2020.
11. 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.
12. Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on AI*, 2010.
13. 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.
14. 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.
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. Omaima Fallatah, Ziqi Zhang, and Frank Hopfgartner. A gold standard dataset for large knowledge graphs matching. In *Ontology Matching 2020: Proceedings of the 15th International Workshop on Ontology Matching co-located with (ISWC 2020)*, 2020.
30. 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.
31. I. Harrow et al. Ontology mapping for semantically enabled applications. *Drug Discovery Today*, 2019.
32. 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.
33. 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.
34. Sven Hertling and Heiko Paulheim. Dbkwik: A consolidated knowledge graph from thousands of wikis. In *Proceedings of the International Conference on Big Knowledge*, 2018.
35. Sven Hertling and Heiko Paulheim. Dbkwik: extracting and integrating knowledge from thousands of wikis. *Knowledge and Information Systems*, 2019.
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. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. Christian Meilicke. *Alignment Incoherence in Ontology Matching*. PhD thesis, University Mannheim, 2011.
49. 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.
50. Franck Michel, Olivier Gargominy, Sandrine Tercerie, and Catherine Faron-Zucker. A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF. In Alsayed Algergawy, Naouel Karam, Friederike Klan, and Clément Jonquet, editors, *Proceedings of the 2nd International Workshop on Semantics for Biodiversity co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22nd, 2017*, volume 1933 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.
51. Boris Motik, Rob Shearer, and Ian Horrocks. Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009.
52. 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.
53. 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.
54. Jan Portisch, Sven Hertling, and Heiko Paulheim. Visual analysis of ontology matching results with the melt dashboard. In *European Semantic Web Conference*, pages 186–190, 2020.
55. Mina Abd Nikooie Pour, Alsayed Algergawy, Reihaneh Amini, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Clément Jonquet, Naouel Karam, Abderrahmane Khiat, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Pascal Hitzler, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Élodie Thiéblin, Cássia Trojahn, Jana Vatascinová, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2020. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 2, 2020*, volume 2788 of *CEUR Workshop Proceedings*, pages 92–138. CEUR-WS.org, 2020.

56. 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.
57. 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.
58. 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.
59. 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.
60. Alessandro Solimando, Ernesto Jimenez-Ruiz, and Giovanna Guerrini. Minimizing conservativity violations in ontology alignments: Algorithms and evaluation. *Knowledge and Information Systems*, 2016.
61. Christian Strobl. *Encyclopedia of GIS*, chapter Dimensionally Extended Nine-Intersection Model (DE-9IM), pages 240–245. Springer, 2008.
62. Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, 2007.
63. York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the Workshop on Evaluation of Ontology-based Tools (EON), Hiroshima (JP)*, 2004.
64. 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.
65. 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.
66. 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.
67. 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,
Trento, Toulouse, Prague, Manhattan, Dublin
December 2021