Results of the Ontology Alignment Evaluation Initiative 2022

Mina Abd Nikooie Pour[1], Alsayed Algergawy[2], Patrice Buche[3], Leyla J. Castro[4], Jiaoyan Chen[5], Hang Dong[6], Omaima Fallatah[7], Daniel Faria[8], Irini Fundulaki[9], Sven Hertling[10], Yuan He[6], Ian Horrocks[6], Martin Huschka[11], Liliana Ibanescu[12], Ernesto Jiménez-Ruiz[13], Naouel Karam[14], Amir Laadhar[15], Patrick Lambrix[1,16], Huanyu Li[1], Ying Li[1], Franck Michel[17], Engy Nasr[18], Heiko Paulheim[10], Catia Pesquita[19], Tzanina Saveta[9], Pavel Shvaiko[20], Cassia Trojahn[21], Chantelle Verhey[22], Mingfang Wu[23], Beyza Yaman[24], Ondrej Zamazal[25] and Lu Zhou[26]

[1]*Linköping University & Swedish e-Science Research Center, Linköping, Sweden*
[2]*Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Germany*
[3]*UMR IATE, INRAE, University of Montpellier, France*
[4]*ZB MED Information Centre for Life Sciences, Germany*
[5]*Department of Computer Science, The University of Manchester, UK*
[6]*Department of Computer Science, University of Oxford, UK*
[7]*Information School, The University of Sheffield, Sheffield, UK*
[8]*University of Lisbon, Portugal*
[9]*Institute of Computer Science-FORTH, Heraklion, Greece*
[10]*Data and Web Science Group, University of Mannheim, Germany*
[11]*Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institut, EMI, Germany*
[12]*Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA Paris-Saclay, France*
[13]*City, University of London, UK & SIRIUS, University of Oslo, Norway*
[14]*Fraunhofer FOKUS & Institute for Applied Informatics, University of Leipzig, Germany*
[15]*University of Stuttgart, Germany*
[16]*University of Gävle, Sweden*
[17]*University Côte d'Azur, CNRS, Inria*
[18]*Albert Ludwig University of Freiburg, Germany*
[19]*LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[20]*Trentino Digitale SpA, Trento, Italy*
[21]*Institut de Recherche en Informatique de Toulouse, France*
[22]*World Data System, International Technology Office, USA*
[23]*Australian Research Data Commons*
[24]*ADAPT Centre, Trinity College Dublin*
[25]*Prague University of Economics and Business, Czech Republic*
[26]*TigerGraph, Inc. USA*

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. The OAEI 2022 campaign offered 14 tracks and was attended by 18 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 ontology matching systems [1, 2], and which has been run for eighteen 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 addressing 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) [3]. Then, a unique OAEI campaign occurred in 2005 at the workshop on Integrating Ontologies held in conjunction with the International Conference on Knowledge Capture (K-Cap) [4]. From 2006 until the present, the OAEI campaigns were held at the Ontology Matching workshop, collocated with ISWC [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20], which this year took place virtually[2].

From 2011, we have been using an environment for automatically processing evaluations 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 [21] 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 2022 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.

OM-2022: Proceedings of the 17th International Workshop on Ontology Matching, October 2022, Hangzhou, China (Virtual)

CEUR Workshop Proceedings (CEUR-WS.org)

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

[2] http://om2022.ontologymatching.org

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

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 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 [22].

The **MELT framework**[5] [21] 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. The evaluation client[6] allows organizers to evaluate packaged systems whereby multiple submission formats are supported (SEALS packages or matchers implemented as Web services).

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. With the increasing usage of other programming languages than Java and increasing hardware requirements for matching systems, since 2021 the MELT Web interface was introduced in order to address this issue. 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. In 2022, 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.

[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

2.3. OAEI campaign phases

As in previous years, the OAEI 2022 campaign was divided into three phases: preparatory, execution, and evaluation.

In the **preparatory phase**, the test cases were provided to participants in an initial assessment period between June 30^{th} and July 31^{st}, 2022. 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 31^{th}. The execution phase was terminated on September 30^{th}, 2022, 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 14 tracks, all of them including OWL ontologies while only one also including SKOS thesauri, namely the Biodiversity and the Ecology track. 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.

test	formalism	relations	confidence	modalities	language	SEALS	HOBBIT	MELT
T-Box/Schema matching								
anatomy	OWL	=	[0 1]	open	EN	√		√
conference	OWL	=, <=	[0 1]	open+blind	EN			√
multifarm	OWL	=	[0 1]	open+blind	AR, CZ, CN, DE, EN, ES, FR, IT, NL, RU, PT			√
complex	OWL	=	[0 1]	open+blind	EN, ES			√
food	OWL	=	[0 1]	open	EN			√
interactive	OWL	=, <=	[0 1]	open	EN	√		
bio-ML	OWL	=, <=	[0 1]	open	EN			√
biodiv	OWL/SKOS	=	[0 1]	open	EN			√
mse	OWL	=, <=, >=	[0 1]	open	EN			√
crosswalks	OWL	=	[0 1]	open	EN			√
common knowl. graph	OWL	=	[0 1]	open	EN			√
Instance and schema matching								
knowledge graph	OWL	=	[0 1]	open	EN			√
Instance matching or link discovery								
spimbench	OWL	=	[0 1]	open+blind	EN		√	
link discovery	OWL	=	[0 1]	open	EN		√	

Table 1
Tracks in OAEI 2022.

3.1. Anatomy

The anatomy track comprises a single test case consisting of matching two fragments of biomedical ontologies which describe the human anatomy[8] (3304 classes) and the anatomy of the mouse[9] (2744 classes). The evaluation is based on a manually curated reference alignment. This dataset has been used since 2007 with some improvements over the years [23].

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. 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 [24]. This year we

[8] www.cancer.gov/cancertopics/cancerlibrary/terminologyresources
[9] http://www.informatics.jax.org/searches/AMA_form.shtml

again run 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 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.

3.3. Multifarm

The multifarm track [25] 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.

3.4. 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 **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.

This year, the other complex sub-tracks (**Hydrography**, **GeoLink**, **Populated GeoLink** and **Populated Enslaved** datasets) have been discontinued.

3.5. Food

The Food Nutritional Composition track aims at finding alignments between food concepts from CIQUAL[10], the French food nutritional composition database, and food concepts from SIREN[11], the Scientific Information and Retrieval Exchange Network of the US Food and Drug administration. Foods from both databases are described in LanguaL[12], a well-known multilingual thesaurus using faceted classification. LanguaL stands for "Langua aLimentaria" or "language of food" and more than 40,000 foods used in food composition databases are described using LanguaL.

In [26] we propose the method to provide OWL modelling of food concepts from both datasets, CIQUAL[13] and SIREN [14], and a gold standard.

The evaluation was performed using the MELT platform. Every participating system was executed in its standard setting and we compare precision, recall and F-measure as well as the computation time.

3.6. Interactive Matching

The interactive matching track aims to assess the performance of semi-automated matching systems by simulating user interaction [27, 28, 29]. 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 [30, 28].

[10] https://ciqual.anses.fr/
[11] http://langual.org/langual_indexed_datasets.asp
[12] https://www.langual.org/default.asp
[13] https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.15454/6CEYU3
[14] https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.15454/5LLGVY

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

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

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.

3.7. Bio-ML

The Bio-ML track [31] incorporates both equivalence and subsumption ontology matching (OM) tasks for biomedical ontologies, with ground truth (equivalence) mappings extracted from Mondo [32] and UMLS [33] (see Table 2). Mondo aims to integrate disease concepts worldwide, while UMLS is a meta-thesaurus for the biomedical domain. Based on techniques (ontology pruning, subsumption mapping construction, negative candidate mapping generation, etc.) proposed in [31], we introduced five OM pairs with their information reported in Table 3. Each OM pair is accompanied with both equivalence and subsumption matching tasks; each matching task has two data split settings: *(i)* unsupervised setting (90% of the mappings for testing, and 10% for validation), and *(ii)* semi-supervised setting (70% of the mappings for testing, 20% for training, and 10% for validation).

For evaluation, in [31] we proposed both *global matching* and *local ranking*; the former aims to evaluate the overall performance by computing Precision, Recall, and F1 metrics for the output mappings against the reference mappings, while the latter aims to evaluate the ability of distinguishing the correct mapping out of several challenging negatives by ranking metrics Hits@K and MRR. Note that subsumption mappings are inherently incomplete, so only local ranking evaluation is applied for subsumption matching.

[15]Created from OMIM texts by Mondo's pipeline tool avaiable at: https://github.com/monarch-initiative/omim.
[16]Created by the official snomed-owl-toolkit available at: https://github.com/IHTSDO/snomed-owl-toolkit.

Table 2
Information of the source ontologies used for creating the OM datasets in Bio-ML.

Mapping Source	Ontology	Ontology Source & Version	#Classes
Mondo	OMIM	Mondo[15]	44,729
	ORDO	BioPortal, V3.2	14,886
	NCIT	BioPortal, V18.05d	140,144
	DOID	BioPortal, 2017-11-28	12,498
UMLS	SNOMED	UMLS, US.2021.09.01[16]	358,222
	FMA	BioPortal, V4.14.0	104,523
	NCIT	BioPortal, V21.02d	163,842

Table 3
Information of each OM dataset in Bio-ML, where the numbers of equivalence and subsumption reference mappings are reported in **#Refs(≡)** and **#Refs (⊑)**, respectively.

Mapping Source	Ontology Pair	Category	#Refs (≡)	#Refs (⊑)
Mondo	OMIM-ORDO	Disease	3,721	103
	NCIT-DOID	Disease	4,684	3,339
UMLS	SNOMED-FMA	Body	7,256	5,506
	SNOMED-NCIT	Pharm	5,803	4,225
	SNOMED-NCIT	Neoplas	3,804	213

As Bio-ML is a new track in the OAEI this year and it attempts to support machine learning-based OM systems, we adopted a flexible way for evaluating participating systems. First, participants can freely choose any tasks and settings they would like to attend. Second, for systems that have been well-adapted to the MELT platform, we used MELT to produce the output mappings. Third, for systems that have been implemented elsewhere and not easy to be made compatible with MELT, we used their source code. Fourth, we also allowed participants (with trust) to directly upload output mappings if their systems had not been published and had not been made compatible with MELT. All our evaluations were conducted with the DeepOnto[17] library on a local machine with Intel Xeon Bronze 3204 CPU 1.90GHz x11 processors, 126GB RAM, and two Quadro RTX 8000 GPUs. The GPUs were mainly used for training systems that involve deep neural networks.

3.8. Biodiversity and Ecology

The biodiversity and ecology (biodiv) track is motivated by the GFBio[18] (The German Federation for Biological Data) alongside its successor NFDI4Biodiversity[19] and the AquaDiva[20] projects, which aim at providing semantically enriched data management solutions for data capture, annotation, indexing and search [34, 35, 36]. In 2020, we partnered with the D2KAB project[21], which develops the AgroPortal[22] ontology repository, to include new matching tasks involving important

[17] https://krr-oxford.github.io/DeepOnto/#/
[18] www.gfbio.org
[19] www.nfdi4biodiversity.org/en/
[20] www.aquadiva.uni-jena.de
[21] www.d2kab.org
[22] agroportal.lirmm.fr

thesauri 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). In the 2021 edition, we added a task to align between two biological taxonomies with rather different but complementary scopes: the well-known NCBI taxonomy (NCBITAXON), and TAXREF-LD [37]. No matching system was able to achieve this matching task due to the large size of the considered taxonomies. To cope with this issue in this years edition, we split the large matching task into a set of smaller, more manageable subtasks through the use of modularization. We obtained six groups corresponding to the kingdoms: Animalia, Bacteria, Chromista, Fungi, Plantae and Protozoa, leading to six well balanced matching subtasks. Table 4 presents detailed information about the ontologies and thesauri used in this year edition.

Table 4
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 Animalia	OWL	2021-02-15	74729	-
TAXREF-LD Animalia	OWL	2020-06-23 (v13.0)	73528	-
NCBITAXON Bacteria	OWL	2021-02-15	326	-
TAXREF-LD Bacteria	OWL	2020-06-23 (v13.0)	312	-
NCBITAXON Chromista	OWL	2021-02-15	2344	-
TAXREF-LD Chromista	OWL	2020-06-23 (v13.0)	2290	-
NCBITAXON Fungi	OWL	2021-02-15	13149	-
TAXREF-LD Fungi	OWL	2020-06-23 (v13.0)	12732	-
NCBITAXON Plantae	OWL	2021-02-15	27013	-
TAXREF-LD Plantae	OWL	2020-06-23 (v13.0)	26302	-
NCBITAXON Protozoa	OWL	2021-02-15	538	-
TAXREF-LD Protozoa	OWL	2020-06-23 (v13.0)	501	-

3.9. Material Sciences and Engineering (MSE)

Data in Material Sciences and Engineering (MSE) can be characterised by scarcity, complexity and presence of gaps. Therefore the MSE community aims for ontology-based data integration via decentralized data management architectures. Several actors using different ontologies results in the growing demand for automatic alignment of ontologies in the MSE domain.

The MSE track uses small to mid-sized ontologies common in the MSE field that are implemented with and without upper-level ontologies. The ontologies follow heterogeneous design

principles with only partial overlap to each other. The current version v1.1[23] of the MSE track includes three test cases summarised in Table 5, where each test case consists of two MSE ontologies to be matched [***O1*** ; ***O2***] as well as one manual reference alignment ***R*** that can be used for evaluation of the matching task. The benchmark also provides background knowledge ***resource***s.

Table 5
The building blocks of the MSE track (MSE benchmark v1.1).

Inputs	First Test Case	Second Test Case	Third Test Case
O1	Reduced MaterialInformation	MaterialInformation	MaterialInformation
O2	MatOnto	MatOnto	EMMO
R	= , $\subset$, $\supset$ corresp.	= corresp.	= corresp.
resources	Chemical Elements Dictionary (DICT), EMMO		

The MSE track makes use of three different MSE ontologies in total, in each of which an ontology using an upper-level ontology is matched to one without an upper-level. The MaterialInformation[38] domain ontology was designed without upper-level ontology and serves as infrastructure for material information and knowledge exchange (545 classes, 98 properties and 411 individuals). Three out of eight submodules of the MaterialInformation were merged to create the Reduced MaterialInformation (32 classes, 43 properties and 17 individuals) for a more efficient creation of the manual reference alignment in the First Test Case, see Table 5.The MatOnto Ontology v2.1[24] (847 classes, 96 properties and 131 individuals) bases on the upper-level ontology bfo2[25]. The Elementary Multiperspective Material Ontology (EMMO v1.0.0-alpha2)[26], is a standard representational ontology framework based on current materials modelling and characterisation knowledge incorporating an upper-, mid- and domain-level (451 classes, 35 properties). For every test case, a manual reference alignment ***R*** was created in close cooperation with MSE domain experts.

The evaluation was performed using the MELT platform on a Windows 10 system with Intel Core i7 870 CPU @2.93GHz x4 and 16 GB RAM. For the time being, no background knowledge was used for evaluation. Every participating system was executed in its standard setting and precision, recall and F-measure as well as the computation time is compared.

3.10. Crosswalks Data Schema Matching

This is a new track introduced this year. It aims at evaluating the ability of systems to deal with the schema metadata matching task, in particular, with a collection of crosswalks from fifteen research data schemas to Schema.org[39, 40]. It is based on the work carried out by the Research Data Alliance (RDA) Research Metadata Schemas Working Group. The collection serve as a reference for data repositories when they develop their crosswalks, as well as an indication of semantic interoperability among the schemas.

[23] https://github.com/EngyNasr/MSE-Benchmark/releases/tag/v1.1
[24] https://raw.githubusercontent.com/iNovexIrad/MatOnto-Ontologies/master/matonto-release.ttl
[25] http://purl.obolibrary.org/obo/bfo/2.0/bfo.owl
[26] https://raw.githubusercontent.com/emmo-repo/EMMO/1.0.0-alpha2/emmo.owl

Table 6
The number of classes and instances in the two common KGs benchmarks

Dataset	#Classes	#Instances
DBpedia	138	631,461
NELL	134	1,184,377
YAGO	304	5,149,594
Wikidata	304	2,158,547

The dataset is composed of 15 source research metadata describing datasets that have been aligned to Schema.org. The source schemas include discipline agnostic schemas Dublin Core, Data Catalogue Vocabulary (DCAT), Data Catalogue Vocabulary - Application Profile (DCAT-AP), Registry Interchange Format - Collections and Services (RIF-CS), DataCite Schema, Dataverse; and discipline schemas ISO19115-1, EOSC/EDMI, Data Tag Suite (DATS), Bioschemas, B2FIND, Data Documentation Initiative (DDI), European Clinical Research Infrastructure Network (ECRIN), Space Physics Archive Search and Extract (SPASE); as well as CodeMeta for software.

This year a subset of the 16 metadata schemas aligned to schema.org has been considered. This subset corresponds to the set of schemas and vocabularies for which an OWL/RDFS serialisation is available. It involves: Data Catalogue Vocabulary (DCAT-v3), Data Catalogue Vocabulary - Application Profile (DCAT-AP), DataCity, Dublin Core (DC), ISO19115-1 schemas (ISO) and RIFCS.

Using as a reference the manually established correspondences, the evaluation here will be based on the well-know measures of precision, recall and F-measure. The systems have been executed on a Ubuntu Linux machine configured with 32GB of RAM running under a Intel Core CPU 2.00GHz x8 processors.

3.11. Common Knowledge Graphs

This track was introduced to OAEI in 2021, and it evaluates the ability of matching systems to match the schema (classes) in large cross-domain knowledge graphs such as DBpedia [41], YAGO [42] and NELL [43]. 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. In addition to the NELL and DBpedia test case, this year we introduced a new test case for matching classes from YAGO and Wikidata [44]. The numbers of entities in the four KG datasets are illustrated in Table 6.

The NELL and DBpedia benchmark [45] was 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-penalizing matches 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. In terms of YAGO and Wikidata *gold standard*, it was originally created [46] and expanded according to OAEI standard as part of [44].

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.

3.12. 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[27] in the course of the DBkWik project [47, 48]. 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 7.

Table 7
Characteristics of the Knowledge Graphs in the Knowledge Graph track, and the sources they were created from.

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

The evaluation is based on reference correspondences at both schema and instance levels. While the schema level correspondences were created by experts, the instance correspondences were extracted from the wiki page itself. Due to the fact that not all inter wiki links on a page represent the same concept a few restrictions were made: 1) only links in sections with a header containing “link” are used, 2) all links are removed where the source page links to more than one concept in another wiki (ensures the alignments are functional), 3) multiple links which point to the same concept are also removed (ensures injectivity), 4) links to disambiguation pages were manually checked and corrected. Since we do not have a correspondence for each instance, class, and property in the graphs, this gold standard is only a *partial gold standard*.

The evaluation was executed on a virtual machine (VM) with 32GB of RAM and 16 vCPUs (2.4 GHz), with Debian 9 operating system and Openjdk version 1.8.0_265. For evaluating all possible submission formats, MELT framework is used. The corresponding code for evaluation can be found on Github[28].

[27] https://www.wikia.com/
[28] https://github.com/dwslab/melt/tree/master/examples/kgEvalCli

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.[29]

3.13. 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 [49] 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[30] 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.

3.14. 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[31] and Spaten [50].

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

[29] http://oaei.ontologymatching.org/2019/results/knowledgegraph/kgBaselineMatchers.zip
[30] Although the files are called Tbox1 and Tbox2, they actually contain a Tbox and an Abox.
[31] https://www.tomtom.com/en_gr/

Model) model [51]. The benchmark generator behind this test case implements all topological relations of DE-9IM between trajectories in the two dimensional space. To the best of our knowledge such a generic benchmark, that takes as input trajectories and checks the performance of linking systems for spatial data does not exist. The focus for the design was (a) on the correct implementation of all the topological relations of the DE-9IM topological model and (b) on producing datasets large enough to stress the systems under test. The supported relations are: *Equals, Disjoint, Touches, Contains/Within, Covers/CoveredBy, Intersects, Crosses, Overlaps*. The test case comprises tasks for all the DE-9IM relations and for LineString/LineString and LineString/Polygon cases, for both TomTom and Spaten datasets, ranging from 200 to 2K instances.

We did not exceed 64 KB per instance due to a limitation of the Silk system[32] 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.

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 18 participating systems. Table 8 lists the participants and the tracks in which they competed. It is worth mentioning that the first-year Bio-ML track has four additional participants (e.g., BERTMap [52] and BERTSubs [53]) that are not listed in Table 8. This is because they need training and validation which are not yet fully supported by the OAEI evaluation platforms, and thus they were tested locally with Bio-ML results reported, but without an OAEI system submission. Some matching systems participated with different variants (Matcha and 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 9. Of the 10 systems participating in the Anatomy track, 8 achieved an F-measure higher than the StringEquiv baseline. Three systems were first time participants (Matcha, ALIOn and SEBMatcher). Long-term participating systems showed few changes in comparison with previous years with respect to alignment quality (precision, recall, F-measure, and recall+), size and run time. The exceptions were ALIN which increased in size (from 1119 to 1159), F-measure (from 0.835 to 0.852), recall (from 0.726 to 0.752) and recall+ (from 0.438 to 0.501), and LogMapBio increased in size (from 1586 to 1596), recall (from 0.914 to 0.919) and recall+ (from 0.773 to 0.787). In terms of run time, 4 out of 10 systems computed an alignment in less than 100 seconds. LogMapLt remains the system with the

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

Table 8
Participants and the status of their submissions.

System	A-LIOn	ALIN	AMD	ATMatcher	CIDER-ML	DLinker	DS-JedAI	GraphMatcher	KGMatcher+	LogMap	LogMap-Bio	LogMapLt	LSMatch	LSMatchMulti	Matcha	SEBMather	TOMATO	WomboCombo	Total=18
anatomy	●	●	●	●	○	○	○	○	○	●	●	●	●	○	●	●	○	○	10
conference	●	●	●	●	○	○	○	●	●	●	○	●	●	○	●	●	●	○	12
multifarm	○	○	○	○	●	○	○	○	○	●	○	●	●	●	○	○	○	○	5
complex	○	○	◐	○	○	○	○	○	○	○	○	○	○	○	◐	○	○	○	2
food	○	○	●	○	○	○	○	○	○	●	○	●	○	○	●	○	○	○	4
interactive	○	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	2
bio-ML	○	○	◐	◐	○	○	○	○	○	◐	○	◐	◐	○	◐	○	○	○	6
biodiv	○	○	○	○	○	○	○	○	○	●	●	●	○	○	◐	○	○	○	4
mse	●	○	◐	○	○	○	○	○	○	●	○	●	○	○	●	○	○	○	4
commonKG	○	○	●	●	○	○	○	○	●	●	○	●	●	○	●	○	○	○	7
crosswalks	○	○	●	○	○	○	○	○	○	●	○	●	○	○	●	○	○	○	4
spimbench	○	○	○	○	○	●	○	○	○	●	○	○	○	○	○	○	○	○	2
link discovery	○	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○	○	2
knowledge graph	○	○	●	●	○	○	○	○	●	●	○	○	●	○	●	○	○	●	7
total	3	3	9	5	3	2	1	1	3	12	2	9	6	1	10	2	1	0	71

shortest runtime. Regarding quality, Matcha achieved the highest F-measure (0.941) and recall+ (0.817), but four other systems obtained an F-measure above 0.88 (SEBMatcher, LogMapBio, LogMap, and AMD) 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. Two systems produced coherent alignments (LogMapBio and LogMap).

4.3. Conference

The conference evaluation results using the sharp reference alignment *rar2* are shown in Table 10. For the sake of brevity, only results with this reference alignment and considering both classes and properties are shown. For more detailed evaluation results, please check conference track's web page.

With regard to two baselines we can group tools according to system's position: six systems outperformed above both baselines (ALIN, ATMatcher, GraphMatcher, LogMap, LogMapLt, and SEBMatcher); two systems performed better than StringEquiv baseline (AMD, LSMatch), and four systems performed worse than both baselines (ALIOn, KGMatcher+, TOMATO, and Matcha). Seven matchers (AMD, ALIN, ALIOn, ATMatcher, KGMatcher+, LSMatch, and SEBMatcher) do not match properties at all. On the other side, Matcha does not match classes at all, while it dominates in matching properties. 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

Table 9
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
Matcha	37	1482	0.951	0.941	0.93	0.817	-
SEBMatcher	35602	1402	0.945	0.908	0.874	0.674	-
LogMapBio	1183	1596	0.873	0.895	0.919	0.787	√
LogMap	9	1402	0.917	0.881	0.848	0.602	√
AMD	160	1299	0.953	0.88	0.817	0.522	-
ALIN	374	1159	0.984	0.852	0.752	0.501	-
LogMapLt	3	1147	0.962	0.828	0.728	0.288	-
ATMatcher	156	1037	0.978	0.794	0.669	0.133	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
LSMatch	20	1009	0.952	0.761	0.634	0.037	-
ALIOn	26134	1913	0.364	0.407	0.46	0.136	-

Table 10
The highest average $F_{[0.5|1|2]}$-measure and their corresponding precision and recall for each matcher with its F_1-optimal threshold (ordered by F_1-measure). Inc.Align. means number of incoherent alignments. Conser.V. means total number of all conservativity principle violations. Consist.V. means total number of all consistency principle violations.

System	Prec.	$F_{0.5}$-m.	F_1-m.	F_2-m.	Rec.	Inc.Align.	Conser.V.	Consist.V.
LogMap	0.76	0.71	0.64	0.59	0.56	0	21	0
GraphMatcher	0.75	0.7	0.63	0.58	0.55	6	21	61
SEBMatcher	0.79	0.7	0.6	0.52	0.48	4	6	50
ATMatcher	0.69	0.64	0.59	0.54	0.51	1	72	8
ALIN	0.82	0.7	0.57	0.48	0.44	0	2	0
LogMapLt	0.68	0.62	0.56	0.5	0.47	3	97	18
edna	0.74	0.66	0.56	0.49	0.45			
AMD	0.82	0.68	0.55	0.46	0.41	1	2	6
LSMatch	0.83	0.69	0.55	0.46	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
ALIOn	0.66	0.44	0.3	0.22	0.19	3	17	49
TOMATO	0.09	0.11	0.16	0.28	0.6	15	4	777
Matcha	0.37	0.2	0.12	0.08	0.07	2	3	24

plotted in Figure 1. Systems are represented as squares or triangles, whereas the baselines are represented as circles.

The Conference evaluation results using the *uncertain reference alignments* are presented in Table 11. Out of the 12 alignment systems, 8 (ALIN, ALION, AMD, KGMatcher+, LogMapLt, LSMatch, SEBMatcher, TOMATO) use 1.0 as the confidence value for all matches they identify. The remaining 4 systems (ATMatcher, GraphMatcher, LogMap, Matcha) have a wide variation

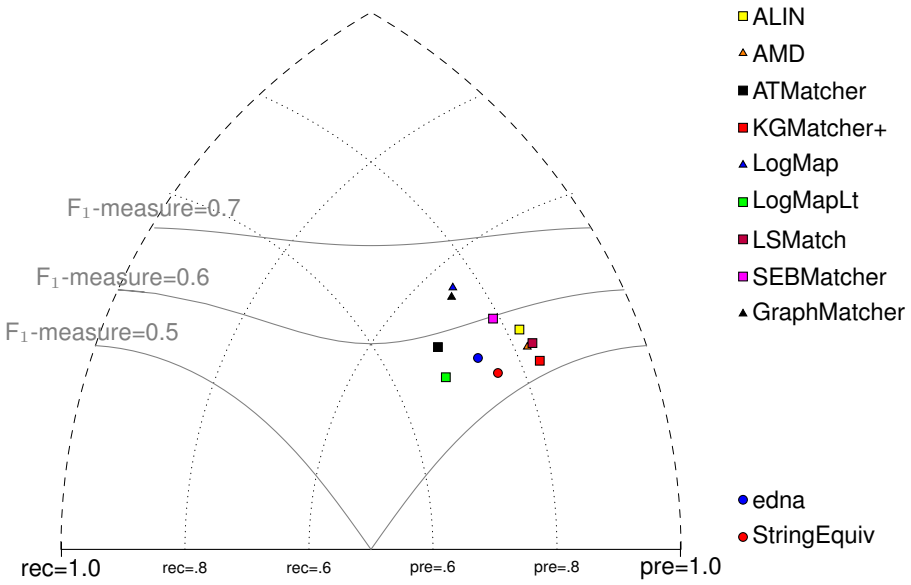

Figure 1: Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F_1-measure are depicted by areas bordered by corresponding lines F_1-measure=0.[5|6|7].

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

	Sharp			Discrete			Continuous		
System	Prec	F-ms	Rec	Prec	F-ms	Rec	Prec	F-ms	Rec
ALIN	0.88	0.61	0.47	0.88	0.70	0.59	0.87	0.71	0.60
ALION	0.75	0.34	0.22	0.75	0.40	0.27	0.75	0.41	0.28
AMD	0.87	0.58	0.43	0.87	0.66	0.53	0.86	0.67	0.55
ATMatcher	0.74	0.62	0.53	0.77	0.67	0.59	0.76	0.68	0.62
GraphMatcher	0.80	0.67	0.57	0.76	0.72	0.68	0.75	0.72	0.68
KGMatcher+	0.88	0.55	0.40	0.88	0.64	0.50	0.88	0.65	0.51
LogMap	0.81	0.68	0.58	0.81	0.70	0.62	0.80	0.66	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	0.87	0.66	0.53	0.88	0.67	0.54
Matcha	0.38	0.13	0.08	0.35	0.14	0.09	0.35	0.12	0.08
SEBMatcher	0.84	0.63	0.50	0.81	0.70	0.61	0.81	0.71	0.62
TOMATO	0.09	0.16	0.63	0.08	0.15	0.74	0.08	0.15	0.73

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 performed the same or better in terms of F-measure. Changes in F-measure of discrete cases ranged from 3 to 18 percent over the sharp reference alignment. ALION is the system whose performance surges

Table 12
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
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
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
LogMapLt	0	0.23	0.26	0.34	0.48	0.67
Matcha	0.8	0.07	0.08	0.09	0.11	0.13

most (18%), followed by KGMatcher+ (16%) and LSMatch (16%). 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. GraphMatcher produces the highest F-measure under both the continuous (72%) and discrete (72%) evaluation methodologies, indicating that this system's confidence evaluation does a good job of reflecting cohesion among experts on this task. Of the remaining systems, LogMap has relatively small drops in F-measure when moving from discrete to continuous evaluation, while Matcha drops 14 percent in F-measure.

Overall, in comparison with last year, the F-measures of most returning matching systems essentially held constant when evaluated against the uncertain reference alignments. ALIN, ALION, GraphMatcher, Matcha, SEBMatcher are 6 new systems participating in this year. ALION's performance increases 18 percent in discrete case and 20 percent in continuous case in terms of F-measure over the sharp reference alignment from 0.34 to 0.40 and 0.41 respectively, which it is mainly driven by increased recall. ALIN, GraphMatcher, and SEBMatcher also perform significantly 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, Matcha outputs low precision and recall among three different versions of reference alignment in general because it assigns the threshold to zero and the matches with relatively high confidence value even the labels of two entities have low string similarity, for example, "hasBid" and "hasPart" has similarity over 0.63 and "addedBy" and "awarded_by" also have similarity over 0.66. Reasonably, it achieves slightly better recall from sharp to discrete case (13%), but the precision and F-measure both drop slightly. TOMATO returns better recall in both discrete and continuous cases. but the precision is significantly lower that other systems, because it outputs multiple matches for same entity and assigns the confidence value as 1.0.

This year we again conducted experiment of matching *cross-domain DBpedia ontology* to three OntoFarm ontologies. The DBpedia ontology has been filtered to the dbpedia namespace

Table 13
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.
CIDER-LM	157	.16	.25	.58
LSMatch	33	.24	.038	.21
LSMatch Multilingual	69	.68	.47	.36
LogMap	9	.72	.44	.31
LogMapLt	175	.24	.038	.02

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 [54]. Out of 12 systems 6 (ATMatcher, KGMatcher+, LogMap, LogMapLt, LSMatch, and Matcha) managed to match DBpedia to OntoFarm ontologies.

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

We can see four systems (LogMap, ATMatcher, KGMatcher+, and LSMatch) perform better than two baselines. LogMap dominates with 0.61 of F1-measure. Most systems achieve lower scores of measures than in the case of matching domain ontologies except KGMatcher+. This shows that these test cases are more difficult for traditional ontology matching systems.

4.4. Multifarm

This year, 5 systems have registered to participate in the MultiFarm track: CIDER-LM, LSMatch, LSMatch Multilingual, LogMap and LogMapLt. The number of participating tools is stable with respect to the last 4 campaigns (6 in 2021, 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 ALOD2Vec, AML, ATMatcher and Wiktionary. But we received new participation from CIDER-LM, LSMatch and LSMatch Multilingual. The reader can refer to the OAEI papers for a detailed description of the strategies adopted by each system.

The Multifarm evaluation results based on the blind dataset are presented in Table 13 demonstrating the aggregated results for the matching tasks. They have been computed using the MELT framework without applying any threshold to 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). 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.

The systems have been executed on a Ubuntu Linux machine configured with 32GB of RAM running under an Intel Core CPU 2.00GHz x8 processors. All measurements are based on a single run. As for each campaign, we observed large differences in the time required for a system

Table 14
Food results per matcher. Time is measured in seconds.

System	Corresp.	Precision	Recall	F1-measure	Time(s)
AMD	0				
LogMap	71	0.0779	0.0822	0.0805	8
LogMapLight	658	0	0	0	14
Matcha	288	0.0833	0,3287	0,13296	38

to complete the 55 x 24 matching tasks: CIDER-LM (157 minutes), LSMatch (33 min), LSMatch Multilingual (69 min), LogMap (9 minutes) and LogMapLt (175 minutes). When we compare the times to the last year's campaign, we can see that LogMap has a stable 9 min execution whereas LogMapLt improved the timing from 212 min to 175 min. Since the other tools are participating for the first time their timings are not comparable. These measurements are only indicative of the time the systems required for finishing the task in a common environment. LSMatch Multilingual outperformed all other systems in terms of F-measure (0.47) whereas CIDER-LM outperformed all other systems in terms of Recall (0.58) and LogMap outperformed all other systems in terms of Precision (0.72).

It is seen that a similar number of systems are participating in the campaign through the years. However, there is a dynamicity of the tools, such that, each year participating tools vary. In 2022, we had 7 systems participating in the campaign where 5 of them were new systems and 2 of them were long-term participating systems. As observed in several campaigns, still, all systems privilege precision in detriment to recall (recall below 0.50) and the results are below the ones obtained for the Conference original dataset.

4.5. Complex Matching

Unfortunately, this track is not attracting many participants since last year. This year, only MatchaC and AMD (for some complex subtracks) have been registered to participate. We lost AMLC, AROA and CANARD with a newcomer MatchaC.

The **Conference subtrack** of the complex track had only one participant MatchaC. However, MatchaC failed to generate alignments. The **Hydrography**, **GeoLink**, and **Populated Enslaved** datasets, as introduced before, have been also discontinued after the announcement of the datasets. For the **Taxon** dataset, as for the Conference dataset, MatchaC and AMD failed to generate alignments. Contrary to last year, we did not run this year the systems generating simple alignments, as simple alignments for this task are usually rather obvious.

4.6. Food

This is the first year of the track and four systems were registered: AMD, LogMap, LogMapLite and Matcha. The evaluation results are presented in Table 14.

The test case evaluates matching systems regarding their capability to find "equal" (=), correspondences between the CIQUAL ontology and the SIREN ontology. None of the evaluated systems finds correspondences other than "equal" (=). All evaluated systems compute the alignment in less than a minute. LogMapLight stands out for its high number of correspondences and

performance indicators all equal to zero. LogMap stands out for its very fast calculation time of 8s. LogMap and Matcha have similar results for precision. However, LogMap's recall is 4 times less than Matcha's one. Matcha is the best performing participant in the food test case in terms of precision, recall and F1-measure.

4.7. Interactive matching

This year, two systems (ALIN, and LogMap) participated in the Interactive matching track. Their results are shown in Table 15 and Figure 2 for both the Anatomy and Conference datasets.

Table 15
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	0.987	0.92	0.952	0.787	0.987	0.92	0.952	579	1453	1.0	1.0
	0.1	0.91	0.9	0.905	0.754	0.987	0.921	0.953	551	1383	0.661	0.976
	0.2	0.847	0.883	0.865	0.727	0.988	0.924	0.955	536	1350	0.472	0.947
	0.3	0.793	0.865	0.827	0.698	0.988	0.925	0.956	529	1325	0.346	0.914
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	0.893	0.565	0.971	0.803	0.879	388	1164	0.749	0.965
	0.2	0.949	0.823	0.882	0.553	0.948	0.761	0.844	388	1164	0.564	0.925
	0.3	0.938	0.817	0.873	0.543	0.93	0.727	0.816	388	1164	0.439	0.88
Conference Dataset												
ALIN	NI	0.874	0.456	0.599	–	–	–	–	–	–	–	–
	0.0	0.919	0.744	0.822	–	0.919	0.744	0.822	309	815	1.0	1.0
	0.1	0.704	0.706	0.705	–	0.935	0.775	0.847	300	787	0.507	0.991
	0.2	0.569	0.663	0.612	–	0.944	0.796	0.863	291	764	0.307	0.972
	0.3	0.476	0.636	0.545	–	0.951	0.812	0.876	283	741	0.209	0.958
LogMap	NI	0.801	0.58	0.67	–	–	–	–	–	–	–	–
	0.0	0.886	0.61	0.723	–	0.886	0.61	0.723	82	246	1.0	1.0
	0.1	0.852	0.598	0.703	–	0.861	0.574	0.689	82	246	0.688	0.978
	0.2	0.81	0.584	0.679	–	0.828	0.546	0.658	82	246	0.494	0.94
	0.3	0.799	0.587	0.677	–	0.804	0.516	0.629	82	246	0.366	0.901

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

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 makes 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 requests feedback on only selected correspondences candidates (based on their similarity patterns or their involvement in unsatisfiabilities). ALIN and LogMap can both ask the oracle to analyze several conflicting correspondences simultaneously.

The performance of the systems usually improves when interacting with a perfect oracle in comparison with no interaction. ALIN is the system that improves the most, because its high number of oracle requests and its non-interactive performance was the lowest of the interactive systems, and thus the easiest to improve.

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

The impact of the oracle's errors is linear for ALIN 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 [55]: 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

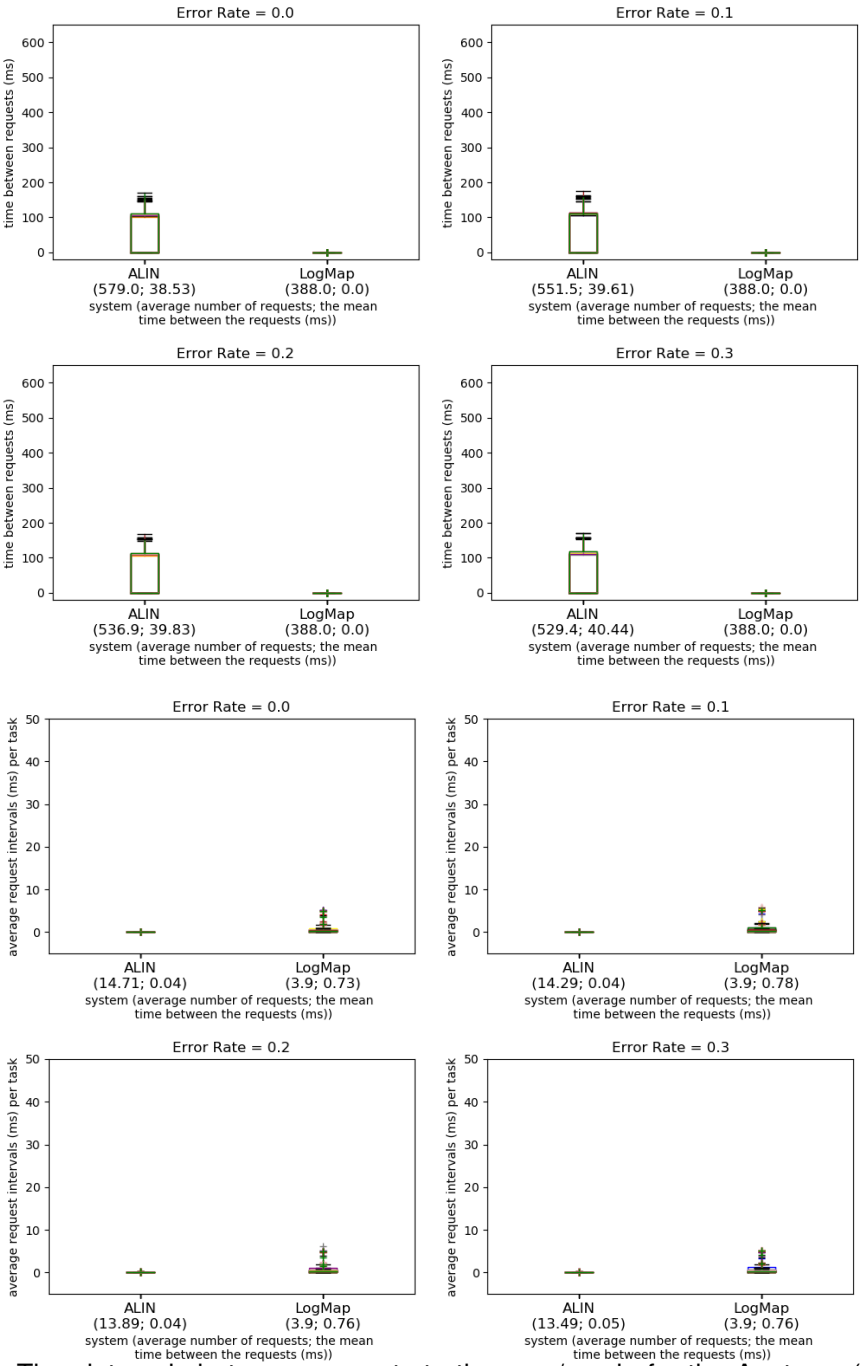

Figure 2: Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

Table 16
Equivalence matching results for OMIM-ORDO (Disease) in Bio-ML track

	Unsupervised (90% Test Mappings)					Semi-supervised (70% Test Mappings)				
System	Precision	Recall	F-score	MRR	Hits@1	Precision	Recall	F-score	MRR	Hits@1
LogMap	0.827	0.498	0.622	0.803	0.742	0.783	0.547	0.644	0.821	0.743
LogMap-Lite	0.935	0.259	0.405	-	-	0.932	0.519	0.667	-	-
AMD	0.664	0.565	0.611	-	-	0.792	0.528	0.633	-	-
BERTMap	0.730	0.572	0.641	0.873	0.817	0.575	0.784	0.664	0.965	0.947
BERTMap-Lite	0.819	0.499	0.620	0.776	0.729	0.775	0.713	0.743	0.900	0.876
Matcha	0.743	0.508	0.604	-	-	0.704	0.564	0.626	-	-
Matcha-DL	-	-	-	-	-	0.956	0.615	0.748	0.654	0.640
ATMatcher	0.940	0.247	0.391	-	-	0.835	0.286	0.426	-	-
LSMatch	0.650	0.221	0.329	-	-	0.877	0.238	0.374	-	-

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 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).

4.8. Bio-ML

Our results include ten tables in total: five tables for equivalence matching and five tables for subsumption matching, where each table corresponds to an OM pair and includes results of both the supervised and semi-supervised settings. See Table 16 for equivalence matching results of OMIM-ORDO (Disease) and Table 17 for subsumption matching results of SNOMED-NCIT (Neoplas). For the full results, please refer to the OAEI 2022 website or the Bio-ML website[33].

Briefly, we have the following participants for equivalence matching: (i) machine learning-based systems including BERTMap [52], BERTMap-Lite [52], AMD [56], Matcha [57] and Matcha-ML [57]; and (ii) traditional systems including LogMap [58], LogMap-Lite, ATMatcher [59], LSMatcher [60]. Note that -Lite means a lightweight version using literal matching. Machine learning-based systems generally perform better with Match-DL attaining best F1 on 4 out of 5 semi-supervised tasks, BERTMap (and -Lite) attaining best F1 on 4 out of 5 unsupervised tasks, and best ranking scores of all tasks. For subsumption matching, all the three participants Word2Vec plus Random Forest (RF), OWL2Vec* [61] plus RF, BERTSubs (IC) [53] BERTSubs performs the best on 2 out of 5 subsumption tasks, while OWL2Vec* performs the best on the remaining 3. Overall, subsumption matching is more challenging than equivalence matching as seen from the lower scores in general. This is intuitive because simple string similarity patterns can play a role in equivalence matching but not likely in subsumption matching.

[33] https://www.cs.ox.ac.uk/isg/projects/ConCur/oaei/2022/

Table 17
Subsumption matching results for SNOMED-NCIT (Neoplas) in Bio-ML track

	Unsupervised (90% Test Mappings)				Semi-supervised (70% Test Mappings)			
System	MRR	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10
Word2Vec+RF	0.512	0.368	0.694	0.834	0.577	0.433	0.773	0.880
OWL2Vec*+RF	0.603	0.461	0.782	0.860	0.666	0.547	0.827	0.880
BERTSubs (IC)	0.530	0.333	0.786	0.948	0.638	0.463	0.859	0.953

4.9. Biodiversity and Ecology

This year, only the LogMap family systems (LogMap, LogMapBio and LogMapLt) alongside Matcha 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 18.

In comparison to the previous year, a smaller number of systems succeeded in generating alignments for the track tasks. The results of the participating systems are comparable to last year in terms of F-measure. In terms of run time, LogMapBio took longer due to the loading of mediating ontologies from BioPortal. Regarding the ENVO-SWEET task, only the LogMap family systems achieved it with a similar performance to last year. The ANAEETHES-GEMET and AGROVOC-NALT matching tasks have the particularity of being resources developed in SKOS. Only LogMapLt could handle the task based on ontology files resulting from an automatic transformation of SKOS files into OWL. For the transformation, we made use of a source code directly derived from the AML ontology parsing module, kindly provided to us by its developers. LogMap, LogMapLt and Matcha performed well on most NCBITAXON-TAXREF-LD subtasks, with slightly the same levels of precision and recall, the larger subtask could only be handled by LogMapLt. Overall, in this fifth evaluation, the number of participating systems decreased considerably and the performance of the successful ones remained similar.

4.10. Material Sciences and Engineering (MSE)

This year five systems registered on the MSE track, each of which was used for evaluation with the three test cases of the MSE benchmarkk. AMD produced errors and an empty alignment file, so results are only available for four of the matchers: A-LIOn, LogMap, LogMapLt, Matcha. The evaluation results are shown in Table 19.

The first test case evaluates matching systems regarding their capability to find ”equal” (=), ”superclass” ($>$) and ”subclass” ($<$) correspondences between the mid-sized MatOnto and the small-sized (since reduced) MaterialInformation ontology. None of the evaluated systems finds correspondences other than ”equal” (=). All evaluated systems compute the alignment in less than a minute. LogMap stands out for its very fast calculation time of 9s and the maximum precision of 1.0. However, since only one correspondence was found by LogMap, the recall and hence the F1-measure is low (0.083). In direct comparison, LogMapLt calculates the alignment in three times the time and achieves much lower precision (0.4) but due to a greater amount of correctly found correspondences the F1-measure is the best of the tested systems in the first test case - although still low with 0.142. A-LIOn finds the highest number of correspondences, but of those 23 found correspondences 20 are false positives which results in the second best F1-measure

Table 18
Results for the Biodiversity & Ecology track.

System	Time (HH:MM:SS)	Number of mappings	Precision	Recall	F-measure
ENVO-SWEET task					
LogMap	00:00:25	676	0.781	0.656	0.713
LogMapBio	01:00:03	697	0.753	0.652	0.699
LogMapLt	00:07:32	576	0.829	0.594	0.692
ANAEETHES-GEMET task					
LogMapLt	00:00:03	182	0.840	0.458	0.593
AGROVOC-NALT task					
LogMapLt	00:00:10	19185	0.744	0.953	0.836
NCBITAXON-TAXREFLD Animalia task					
LogMapLt	00:00:43	72010	0.665	0.993	0.796
NCBITAXON-TAXREFLD Bacteria task					
LogMap	00:00:01	304	0.575	1.0	0.730
LogMapLt	00:00:00	290	0.6	0.994	0.748
Matcha	00:00:05	303	0.577	1.0	0.732
NCBITAXON-TAXREFLD Chromista task					
LogMap	00:00:04	2218	0.623	0.985	0.764
LogMapLt	00:00:01	2165	0.637	0.982	0.773
Matcha	00:00:15	2219	0.623	0.984	0.763
NCBITAXON-TAXREFLD Fungi task					
LogMap	00:00:39	12949	0.783	0.998	0.878
LogMapLt	00:00:07	12929	0.783	0.997	0.877
Matcha	00:00:51	12936	0.785	0.999	0.879
NCBITAXON-TAXREFLD Plantae task					
LogMapLt	00:00:17	26359	0.746	0.987	0.849
Matcha	00:01:15	26675	0.741	0.992	0.848
NCBITAXON-TAXREFLD Protozoa task					
LogMap	00:00:01	496	0.719	1.0	0.837
LogMapLt	00:00:00	477	0.746	0.997	0.853
Matcha	00:00:11	494	0.722	1.0	0.839

at the slowest pace. Matcha finds 4 incorrect crorrespondences thus is the worst performing participant in the MSE test case one. Investigating the alignment produced by Matcha, classes are wrongly matched to object properties, e.g. ”Temperature” =” hasTemperature”.

The second test case evaluates the matching systems to find correspondences between the large-sized MaterialInformation and the mid-sized BFO-based MatOnto. In comparison to the first test case, two of the four evaluated systems (A-LIOn, Matcha) need much longer to calculate the alignment. Surprisingly two of the systems are even quicker (LogMap, LogMapLight) than in the first test case. A-LIOn finds a large number of correspondences, hence has the highest recall

of the evaluated systems but 100 out of the 163 found correspondences are incorrect. This results in a moderate F1-measure of 0.271 and a rather slow calculation time of over 3 minutes. LogMap stands out again for its very fast computation time of only 3s at a high precision of 0.881. Since LogMap found only 59 correct correspondences out of the 302 reference correspondences, the recall is rather low but the F1-measure is still the highest of the tested systems. LogMapLt is almost 30 times slower than LogMap but finds the same amount of correspondences with 2 additional false positives, so it achieves a slightly lower overall F1-measure than LogMap. Matcha finds 6 wrong correspondences where classes are matched to object properties as in the first test case.

The third test case evaluates matching systems to find correspondences between the large-sized MaterialInformation and the mid-sized EMMO. All evaluated systems compute the alignments in under 3 minutes. Surprisingly, A-LIOn takes the longest to compute the alignments but does not find any correspondence which might be due to reasoning errors that were produced for EMMO. LogMap again stands out for the fast computation time and high precision with 53 correct correspondences out of the 56 in total. Although LogMap misses out 10 reference correspondences, the F1-measure of 0.891 is the best of the whole MSE track. LogMapLt is 6 times slower than LogMap with a slightly lower precision and 2 additional false positives. Due to those, LogMapLt achieves a slightly worse F1-measure of 0.857 which is still the second best of the whole MSE track. Matcha finds 2 correct correspondences out of the 63 reference correspondences which results in the only non-zero recall for Matcha in the MSE track along with a fair precision of 0.5 at a rather fast calculation time of 21s.

Table 19
Results for the three test cases of the MSE track.

System	Corresp.	Precision	Recall	F1-Measure	Time [s]
First Test Case					
A-LIOn	23	0.130	0.130	0.130	38
LogMap	1	1.000	0.043	0.083	9
LogMapLt	5	0.400	0.087	0.143	27
Matcha	4	0.000	0.000	0.000	22
Second Test Case					
A-LIOn	163	0.387	0.209	0.271	208
LogMap	67	0.881	0.195	0.320	3
LogMapLt	67	0.851	0.189	0.309	83
Matcha	6	0.000	0.000	0.000	15
Third Test Case					
A-LIOn	0	0.000	0.000	0.000	135
LogMap	56	0.946	0.841	0.891	14
LogMapLt	56	0.911	0.810	0.857	84
Matcha	4	0.500	0.032	0.060	21

In summary, LogMap stands out for its very fast computing speed with very high precision at

the same time. LogMapLt is significantly slower in every test case and almost constantly shows worse results - only in the first test case the recall of LogMapLight is higher than for LogMap. In our opinion, LogMap is definitely recommended for MSE applications where high precision is demanded. In comparison to that, LogMapLight does not appear to bring any decisive advantage over LogMap.

Matcha in its current implementation is not recommended for MSE applications since it matches classes to properties.

A-LIOn produces moderate results but does not bring any advantage over LogMap. Furthermore, A-LIOn produces errors while reasoning on EMMO. The latter is the only one of the MSE ontologies used with a significant proportion of essential axioms. According to the annotations in EMMO, this ontology exclusively can be inferred with the FaCT++ reasoner. That might be a cause for the occurring reasoning errors of A-LIOn and bad results in the third test case.

None of the evaluated matcher finds all reference correspondences correctly and none of the matchers.

4.11. Common Knowledge Graphs

We evaluated all the participating systems that were packaged as SEALS packages or as web services using Docker (even those not registered to participate on this new track). However, not all systems were able to complete the task as some systems finished with an empty alignment file. Here, we include the results of 8 systems that were able to finish the task within the 24 hours time limit with a non-empty alignment file: LogMap, ATMatcher, Matcha, KGMatcher+, LogMapLite, LogMapKG, LsMatch, and AMD.

Table 20 shows the aggregated results on the two datasets for systems that produced non-empty alignment files. The size column indicates the total number of class alignments discovered by each system. While the majority of the systems discovered alignments at both schema and instance levels, we have only evaluated class alignments, as the two gold standard does not include any instance-level ground truth. Further, Not all systems were able to handle the original dataset versions (i.e., those with all annotated instances). In terms of the NELL-DBpedia test case, LogMap, ATMatcher, KGMatcher+, and AMD were able to generate results when applied to the full-size dataset. While on the YAGO-Wikidata dataset, which is large-scale compared to the first dataset, only ATMatcher, KGMatcher+, and AMD were able to generate alignments with the original dataset. Other systems either fail to complete the task within the allocated 24 hours time limit such as LogMapLite, Matcha, and LsMatch, or produce an empty alignment file such as Matcha, LogMap (only on the YagoWikidata dataset). LogMapKG on the other hand tend to only align instances when it is applied to full-size datasets. Similar to 2021 evaluation results, AMD does generate schema alignments but in the wrong format, therefore, they can not be evaluated.

The resulted alignment files from all the participating systems are available to download on the track's result webpage[34]. On the Nell-DBpedia dataset, all systems were able to outperform the basic string matcher, in terms of f-measure, except for LogMapLite. On the YagoWikidata dataset, two systems were not able to outperform the baseline, which are LogMapLite and

[34] https://oaei.ontologymatching.org/2022/results/commonKG/index.html

Table 20
Results for the Common Knowledge Graphs track

Matcher	Size	Precision	Recall	F1 measure	Time	Dataset Size
NELL-DBpedia						
LogMap	105	0.99	0.80	0.88	00:03:17	original
ATMatcher	104	1.00	0.80	0.89	00:03:10	original
Matcha	104	1.00	0.81	0.90	00:01:00	small
KGMatcher+	117	1.00	0.91	0.95	02:43:50	original
LogMapLite	77	1.00	0.60	0.75	00:26:19	small
LogMapKG	104	0.98	0.80	0.88	00:00:00	small
AMD	102	0.00	0.00	0.00	00:00:23	original
LsMatch	101	0.96	0.75	0.84	00:00:52	small
String Baseline	78	1.00	0.60	0.75	00:00:37	original
YAGO-Wikidata						
LogMap	233	1.00	0.76	0.86	00:01:19	small
ATMatcher	233	1.00	0.77	0.87	00:19:04	original
Matcha	243	1.00	0.80	0.89	00:03:18	small
KGMatcher+	253	0.99	0.83	0.91	02:07:59	original
LogMapLite	211	1.00	0.70	0.81	00:48:19	small
LogMapKG	232	1.00	0.76	0.83	00:00:10	small
AMD	125	0.00	0.00	0.00	00:29:04	original
LsMatch	196	0.96	0.63	0.76	00:02:28	small
String Baseline	212	1.00	0.70	0.82	00:00:02	original

LsMatch. Similar to last year, KGMatcher+ outperforms other systems in terms of f-measure. It achieves 0.95 as f-measure on the NELL-DBpedia test case and 0.91 on the YAGO-Wikidata test case. While other systems were not able to improve last year's results, Matcha has improved last year's AML results with 0.90 f-measure on the NELL-DBpedia test case, and 0.89 on the YAGO-Wikidata.

In terms of runtime, Table 20 also presents the run time as HH:MM:SS where we can observe that all matching were able to finish the task in less than 30 minutes except for KGMatcher+ and LogMapLite. Finally, the dataset size column identifies whether the system was able to perform on the original dataset or only on the smaller version.

4.12. Crosswalks Data Schema Matching

For this first version of the track, four systems registered to participate: AMD, LogMap, LogMapLt and Matcha. Table 21 shows the results for the systems that have generated correspondences. AMD was not able to generate any correspondence. LogMap, LogMapLt and Matcha are the only systems able to generated (few) correct correspondences. The generated correspondences involved mostly classes and properties where labels are the same, for instance: https://schema.org/distribution and http://www.w3.org/ns/dcat#distribution.

With respect to the pairs of schemas, the systems generated a higher number of correspondences for DCAT-v3 and RIFCS. LogMapLt is the system that is able to deal with a higher number of matching pairs. None of the participant systems generated outputs for DC.

Table 21
Results for the Crosswalk task.

Matcha			
	correct	output	expected
dcat3	3	17	42
datacity	0	4	34
LogMap			
	correct	output	expected
dcat3	0	12	42
datacity	0	3	34
rifcs	0	11	24
dcat-ap	0	2	34
LogMapLt			
	correct	output	expected
dcat3	3	41	42
datacity	0	4	34
rifcs	0	9	24
dcat-ap	0	4	34
iso	0	2	42

This task mostly deals with properties of metadata schemes. Still, dealing with properties is a challenging task.

This year, as introduced above, we have used the schemes for which a OWL/RDFS serialization is available, as OAEI matching systems are used to the format. However, this does not reflect the reality of the field, as schemas are not usually exposed in such a structured format. This opens the possibility of providing a dedicated task next year.

4.13. Link Discovery

This year the Link Discovery track counted four participants: DS-JedAI, Silk, RADON and DLinker. DLinker participated for the first time.

We divided the Spatial test cases into four suites. In the 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 other two suites (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 3-4 [35].

The detailed results can also be found in HOBBIT git [36]. Silk and GS-JedAI did not participate for COVERED BY and Silk also did not participate for COVERS. DLinker only participated for EQUALS and OVERLAPS tasks and only for LineStrings to LineStrings.

In the SLL suite, RADON has the best performance in most cases except for the *Touches* and *Intersects* relations. DS-JedAI seems to need the most time while Silk has the second best performance. DLinker perfom well regarding *Overlaps* and also *Equals* for Spaten dataset.

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. Dlinker scales pretty well for the *Overlaps* and also *Equals* for Spaten dataset following the performance of SLL suite.

In the SLP suite, in contrast to the first two suites, RADON has the best performance for all relations. Silk has the second best time performance 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. Again, DS-JedAI scales better in large datasets, thus it needs less time than Silk.

Taking into account the executed test cases we can identify the capabilities of the tested systems as well as suggest some improvements. Three of 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 of those systems did not manage to complete some test cases, mostly *Disjoint*. One system, DLinker only participated for *Equals* and *Overlaps* relations and only for Linestrings to Linestrings test cases.

RADON was the only system that successfully addressed all the tasks, and had the best performance for the SLP and LLP suites, but it can be improved for the *Touches* and *Intersects* relations for the SLL and LLL suites. 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 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. DLinker can be improved to the number of the supported relations as well as to the supported geometries.

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.

[35] In order to make the diagrams more comprehensible we have excluded the extreme values.
[36] https://hobbit-project.github.io/OAEI_2022.html

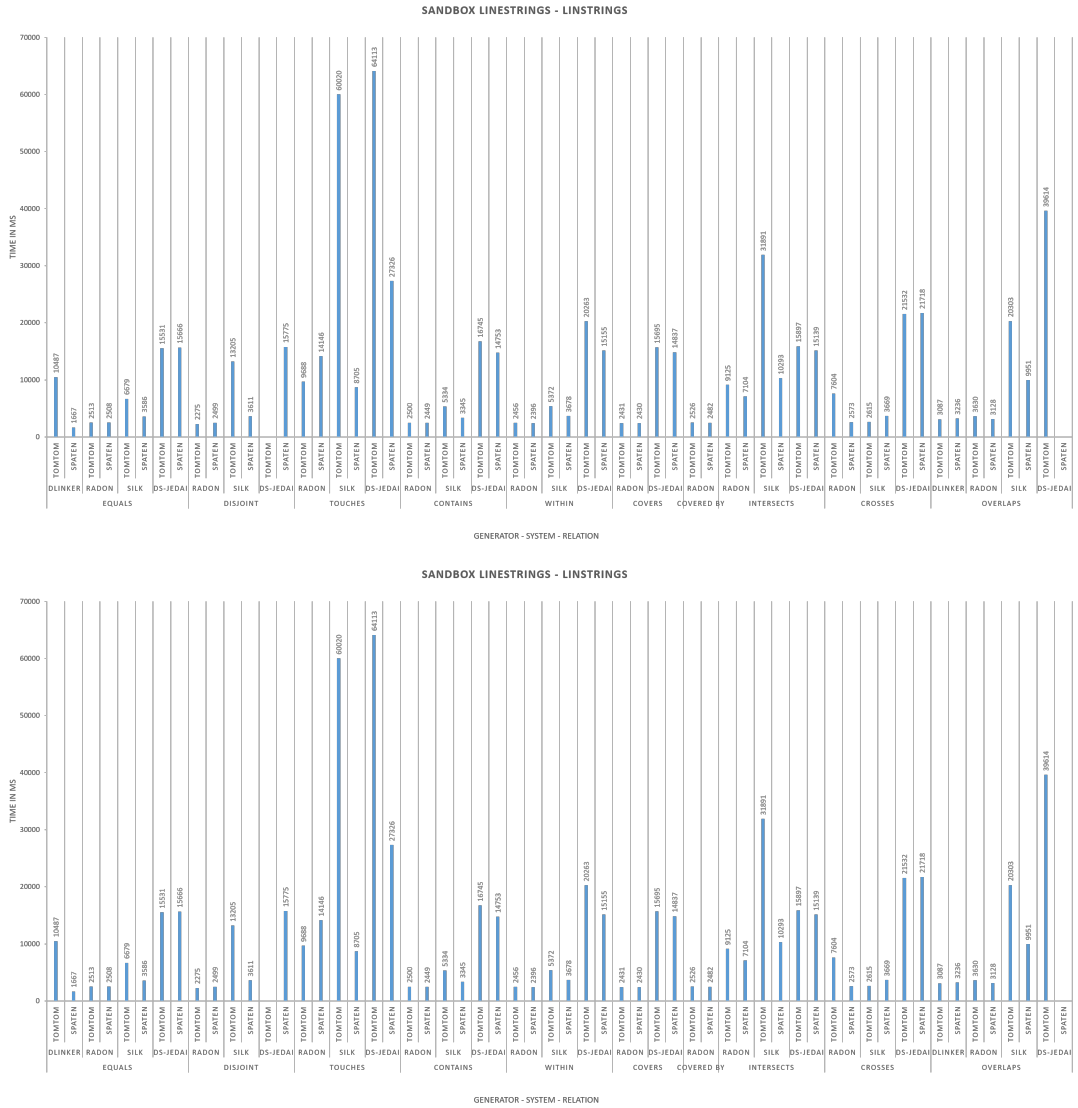

Figure 3: Time performance for TomTom & Spaten SLL (top) and LLL (bottom) suites for DLinker, RADON, Silk and DS-JedAI.

4.14. SPIMBENCH

This year, the SPIMBENCH track counted two participants: LogMap and DLinker. DLinker participated for the first time but only for the Sanbox task while LogMap participates every year. The evaluation results of the track are shown in Table 22. The results can also be found in HOBBIT git [37].

[37] https://hobbit-project.github.io/OAEI_2022.html

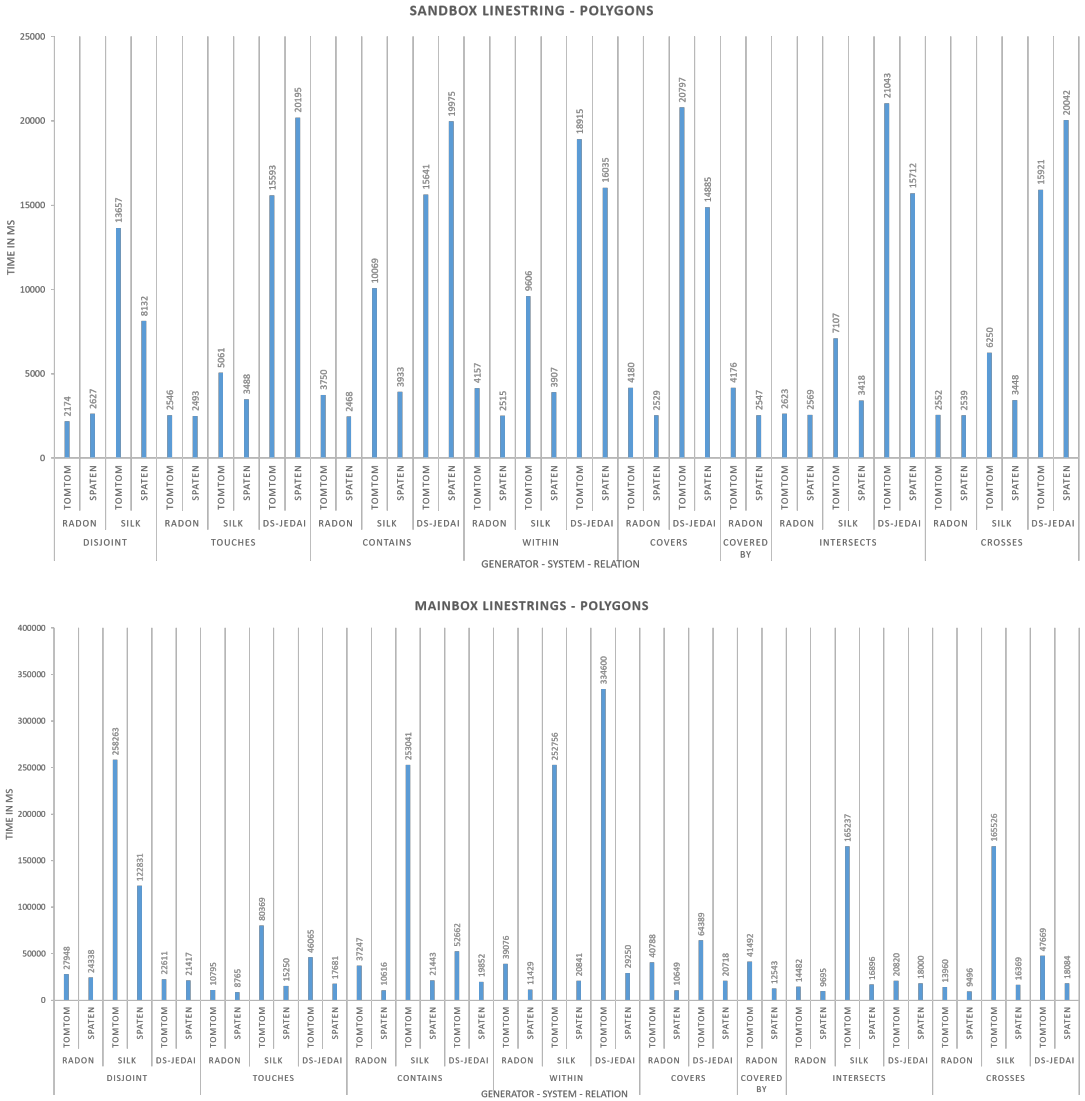

Figure 4: Time performance for TomTom & Spaten SLP (top) and LLP (bottom) suites for RADON, Silk and DS-JedAI.

LogMap has the best performance overall both in terms of F-measure and run time. The run time scaled very well with the increase if the number of instances while we do not have scaling information for DLinker as it did not participate for the large dataset.

4.15. 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

Table 22
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
DLinker	0.7026	0.7907	0.6321	15555
Mainbox Dataset (1800 instances, 50000 triples)				
System	**Fmeasure**	**Precision**	**Recall**	**Time (in ms)**
LogMap	0.7856	0.8801	0.7094	27140

matching task[38] (even those not registered for the track). This revealed that not all systems were able to handle the task, and in the end, 5 matchers can provide results for at least one test case.

Similar to the previous years, some systems like AMD 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 [39].

Table 23 shows the results for all systems divided into class, property, instance, and overall results. This also includes 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, 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.

This years best overall system is ATMatcher. The result of 0.85 is still behind the top result over all years which was 0.87. All other systems, regarding F-Mesasure, are below the baseline which includes the alternative labels. The highest recall is achieved by Matcha (0.88), a new system participating this year. It returns more correspondences than all others (32844.2 on average) but is only able to match instances in this track. AMD returned some correspondences but achieved an overall F-Measure of 0.0 for all test cases. Thus the system is not included in the final table. Detailed results for each test case can be found on the OAEI results page of the track[40].

Property matches are still not created by all systems. KGMatcher, LogMap, and Matcha do not return any of those mappings. One reason might be that the properties are typed as `rdf:Property` and not distinguished into `owl:ObjectProperty` or `owl:DatatypeProperty`. ATMatcher reaches the best score with 0.96 F-Measure. The number of instance correspondences are in the same range (3,641 - 5,872) for all systems except Matcha (32,844) which thus reaches a high recall value.

Regarding runtime, LSMatch (4:17:13) was the slowest system. In comparison to last year with more than 12 hours, the runtimes of this campain is rather good and shows the scalability of the systems. Besides the baselines (which need around 12 minutes for all test cases) ATMatcher (00:18:48) and LogMap (00:55:52) were the fastest systems.

[38] http://oaei.ontologymatching.org/2019/results/knowledgegraph/small_test.zip
[39] http://oaei.ontologymatching.org/2022/results/knowledgegraph/knowledgegraph-alignments.zip
[40] http://oaei.ontologymatching.org/2022/results/knowledgegraph/index.html

Table 23
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						
ATMatcher	00:18:48	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)
KGMatcher	03:01:17	5	21.2	1.00 (1.00)	0.79 (0.79)	0.66 (0.66)
LogMap	00:55:52	5	19.4	0.93 (0.93)	0.81 (0.81)	0.71 (0.71)
LSMatch	04:17:13	5	23.6	0.97 (0.97)	0.78 (0.78)	0.64 (0.64)
Matcha	02:40:21	4	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
property performance						
ATMatcher	00:18:48	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)
KGMatcher	03:01:17	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
LogMap	00:55:52	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
LSMatch	04:17:13	5	85.6	0.73 (0.73)	0.71 (0.71)	0.69 (0.69)
Matcha	02:40:21	4	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
instance performance						
ATMatcher	00:18:48	5	4856.6	0.89 (0.89)	0.84 (0.84)	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)
KGMatcher	03:01:17	5	3789.6	0.94 (0.94)	0.82 (0.82)	0.74 (0.74)
LogMap	00:55:52	5	4012.4	0.90 (0.90)	0.78 (0.78)	0.69 (0.69)
LSMatch	04:17:13	5	5872.2	0.66 (0.66)	0.63 (0.63)	0.60 (0.60)
Matcha	02:40:21	4	32844.2	0.53 (0.66)	0.61 (0.76)	0.72 (0.90)
overall performance						
ATMatcher	00:18:48	5	4961.0	0.89 (0.89)	0.85 (0.85)	0.80 (0.80)
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)
KGMatcher	03:01:17	5	3810.8	0.94 (0.94)	0.82 (0.82)	0.72 (0.72)
LogMap	00:55:52	5	4031.8	0.90 (0.90)	0.77 (0.77)	0.68 (0.68)
LSMatch	04:17:13	5	5981.4	0.66 (0.66)	0.63 (0.63)	0.61 (0.61)
Matcha	02:40:21	4	32844.2	0.53 (0.66)	0.60 (0.76)	0.70 (0.88)

For further analysis of the results, we also provide an online dashboard[41] generated with MELT[62]. It allows to inspect the results on a correspondence level. Due to the large amount of these correspondences, it can take some time to load the full dashboard.

[41] http://oaei.ontologymatching.org/2022/results/knowledgegraph/knowledge_graph_dashboard.html

5. Conclusions and Lessons Learned

In 2022 we witnessed a healthy mix of new and returning systems. Like last year, the distribution of participants by tracks was uneven.

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

According to the **Conference** track there is still need for an improvement with regard to the ability of matching systems to match properties (even, majority systems only match classes; 7 out of 12 systems). Next, much fewer systems, with regard to the last year, managed to match DBpedia ontology to conference ontologies (85% vs. 50%). With respect to the first point, it has been corroborated in the new track (**Crosswalks Data Schema Matching**), that concerns properties in particular. In fact systems still fail in dealing with this kind of ontology entity.

Since the creation of the **Material Sciences and Engineering track** a large amount of new ontologies have been developed and utilized in various MSE applications. In contrast to the early development stages of this track, those ontologies are now easily accessible on the new Matportal[42]. In the future, the MSE track should be updated with the currently most used top and mid-level MSE ontologies, which include the BWMD-mid[43], the MSEO[44] , the PMDco[45], the prov-o[46]. Apart from also considering frequently used domain and application ontologies, also multi-ontology matching, knowledge graph matching and the usage of background knowledge should be considered in this domain where such applications become increasingly important.

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.

In the **Food** track none of the evaluated systems finds all the reference correspondences. The usage of background knowledge available in CIQUAL and SIREN ontologies in terms of food description based on FoodON concepts should be considered in future OAEI campaigns.

The **Bio-ML track** is new to the OAEI. It has several participants for equivalence matching but very few for subsumption matching which is more challenging. The best performing systems are not consistent across tasks and settings, demonstrating the diversity of our datasets. To encourage more participants and provide more convincing results in the future, we will consider creating a new set of OM pairs (different ontologies and/or semantic types from up-to-date sources) and anonymize the testing mappings.

In the **Biodiversity and Ecology track**, none of the systems has been able to detect mappings

[42] https://matportal.org/
[43] https://matportal.org/ontologies/BWMD-MID
[44] https://matportal.org/ontologies/MSEO
[45] https://github.com/materialdigital/core-ontology/
[46] https://www.ebi.ac.uk/ols/ontologies/prov

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. Two systems participated this year. This is puzzling considering that this track is based on the *Anatomy* and *Conference* test cases, and those tracks had 10 and 12 participants, respectively. 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 that still attracts a very few number of participants. This year, no system was able to complete the task. While some sub-tracks have been discontinued, the Taxon track has to evolve, in particular 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**, there is a slight decrease of systems able to solve all test cases. The overall best scores are still unbeaten. Furthermore the proportion of matchers not able to produce property alignments is high. This might change next year with new and improved systems.

This is the second year of the **Common knowledge graphs track**, which challenges matching systems to map the schema of large-scale, automatically constructed, and cross-domain knowledge graphs. The number of participants is similar to last year, i.e. 8 systems. While a number of systems were able to finish the task, other systems still faced problems coping with the scalability issue. Some of the systems were only able to produce alignments when applied to smaller versions of the KGs dataset. Therefore, we still expect those systems to be adapted to the task, and we look forward to having more participants in the next OAEI 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, using SSSOM as alignment format for towards making FAIR alignments, and alternative ways for running resource-consuming systems.

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).

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

Beyza Yaman has been supported by ADAPT SFI Research Centre [grant 13/RC/2106_P2].

Jiaoyan Chen, Hang Dong, Yuan He and Ian Horrocks have been supported by Samsung Research UK (SRUK) and the EPSRC project ConCur (EP/V050869/1).

The Biodiversity and Ecology track has been partially funded by the German Research Foundation in the context of NFDI4BioDiversity 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) 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] J. Euzenat, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, C. Trojahn dos Santos, Ontology alignment evaluation initiative: six years of experience, Journal on Data Semantics XV (2011) 158–192.

[2] J. Euzenat, P. Shvaiko, Ontology matching, 2nd ed., Springer-Verlag, 2013.

[3] Y. Sure, O. Corcho, J. Euzenat, T. Hughes (Eds.), Proceedings of the Workshop on Evaluation of Ontology-based Tools (EON), Hiroshima (JP), 2004.

[4] B. Ashpole, M. Ehrig, J. Euzenat, H. Stuckenschmidt (Eds.), Proc. K-Cap Workshop on Integrating Ontologies, Banff (Canada), 2005. URL: http://ceur-ws.org/Vol-156/.

[5] M. Abd Nikooie Pour, A. Algergawy, F. Amardeilh, R. Amini, O. Fallatah, D. Faria, I. Fundulaki, I. Harrow, S. Hertling, P. Hitzler, M. Huschka, L. Ibanescu, E. Jiménez-Ruiz, N. Karam, A. Laadhar, P. Lambrix, H. Li, Y. Li, F. Michel, E. Nasr, H. Paulheim, C. Pesquita, J. Portisch, C. Roussey, T. Saveta, P. Shvaiko, A. Splendiani, C. Trojahn, J. Vatascinová, B. Yaman, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2021, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn (Eds.), Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 25, 2021, volume 3063 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 62–108. URL: http://ceur-ws.org/Vol-3063/oaei21_paper0.pdf.

[6] M. Abd Nikooie Pour, A. Algergawy, R. Amini, D. Faria, I. Fundulaki, I. Harrow, S. Hertling, E. Jiménez-Ruiz, C. Jonquet, N. Karam, A. Khiat, A. Laadhar, P. Lambrix, H. Li, Y. Li, P. Hitzler, H. Paulheim, C. Pesquita, T. Saveta, P. Shvaiko, A. Splendiani, É. Thiéblin, C. Trojahn, J. Vatascinová, B. Yaman, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2020, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn (Eds.), 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*, CEUR-WS.org, 2020, pp. 92–138. URL: http://ceur-ws.org/Vol-2788/oaei20_paper0.pdf.

[7] A. Algergawy, D. Faria, A. Ferrara, I. Fundulaki, I. Harrow, S. Hertling, E. Jiménez-Ruiz, N. Karam, A. Khiat, P. Lambrix, H. Li, S. Montanelli, H. Paulheim, C. Pesquita, T. Saveta, P. Shvaiko, A. Splendiani, É. Thiéblin, C. Trojahn, J. Vatascinová, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2019, in: Proceedings of the 14th International Workshop on Ontology Matching, Auckland, New Zealand, 2019, pp. 46–85.

[8] A. Algergawy, M. Cheatham, D. Faria, A. Ferrara, I. Fundulaki, I. Harrow, S. Hertling, E. Jiménez-Ruiz, N. Karam, A. Khiat, P. Lambrix, H. Li, S. Montanelli, H. Paulheim, C. Pesquita, T. Saveta, D. Schmidt, P. Shvaiko, A. Splendiani, É. Thiéblin, C. Trojahn, J. Vatascinová, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2018, in: Proceedings of the 13th International Workshop on Ontology Matching, Monterey (CA, US), 2018, pp. 76–116.

[9] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, E. Jiménez-Ruiz, K. Kolthoff, E. Kuss, P. Lambrix, H. Leopold, H. Li, C. Meilicke, M. Mohammadi, S. Montanelli, C. Pesquita, T. Saveta, P. Shvaiko,

A. Splendiani, H. Stuckenschmidt, É. Thiéblin, K. Todorov, C. Trojahn, O. Zamazal, Results of the ontology alignment evaluation initiative 2017, in: Proceedings of the 12th International Workshop on Ontology Matching, Vienna, Austria, 2017, pp. 61–113. URL: http://ceur-ws.org/Vol-2032/oaei17_paper0.pdf.

[10] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, E. Jiménez-Ruiz, E. Kuss, P. Lambrix, H. Leopold, H. Li, C. Meilicke, S. Montanelli, C. Pesquita, T. Saveta, P. Shvaiko, A. Splendiani, H. Stuckenschmidt, K. Todorov, C. Trojahn, O. Zamazal, Results of the ontology alignment evaluation initiative 2016, in: Proceedings of the 11th International Ontology matching workshop, Kobe (JP), 2016, pp. 73–129.

[11] M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, R. Granada, V. Ivanova, E. Jiménez-Ruiz, P. Lambrix, S. Montanelli, C. Pesquita, T. Saveta, P. Shvaiko, A. Solimando, C. Trojahn, O. Zamazal, Results of the ontology alignment evaluation initiative 2015, in: Proceedings of the 10th International Ontology matching workshop, Bethlehem (PA, US), 2015, pp. 60–115.

[12] Z. Dragisic, K. Eckert, J. Euzenat, D. Faria, A. Ferrara, R. Granada, V. Ivanova, E. Jiménez-Ruiz, A. O. Kempf, P. Lambrix, S. Montanelli, H. Paulheim, D. Ritze, P. Shvaiko, A. Solimando, C. T. dos Santos, O. Zamazal, B. C. Grau, Results of the ontology alignment evaluation initiative 2014, in: Proceedings of the 9th International Ontology matching workshop, Riva del Garda (IT), 2014, pp. 61–104. URL: http://ceur-ws.org/Vol-1317/oaei14_paper0.pdf.

[13] B. Cuenca Grau, Z. Dragisic, K. Eckert, J. Euzenat, A. Ferrara, R. Granada, V. Ivanova, E. Jiménez-Ruiz, A. Kempf, P. Lambrix, A. Nikolov, H. Paulheim, D. Ritze, F. Scharffe, P. Shvaiko, C. Trojahn dos Santos, O. Zamazal, Results of the ontology alignment evaluation initiative 2013, in: P. Shvaiko, J. Euzenat, K. Srinivas, M. Mao, E. Jiménez-Ruiz (Eds.), Proceedings of the 8th International Ontology matching workshop, Sydney (NSW, AU), 2013, pp. 61–100. URL: http://oaei.ontologymatching.org/2013/results/oaei2013.pdf.

[14] J. Aguirre, B. Cuenca Grau, K. Eckert, J. Euzenat, A. Ferrara, R. van Hague, L. Hollink, E. Jiménez-Ruiz, C. Meilicke, A. Nikolov, D. Ritze, F. Scharffe, P. Shvaiko, O. Sváb-Zamazal, C. Trojahn, B. Zapilko, Results of the ontology alignment evaluation initiative 2012, in: Proceedings of the 7th International Ontology matching workshop, Boston (MA, US), 2012, pp. 73–115. URL: http://oaei.ontologymatching.org/2012/results/oaei2012.pdf.

[15] J. Euzenat, A. Ferrara, R. van Hague, L. Hollink, C. Meilicke, A. Nikolov, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, C. Trojahn dos Santos, Results of the ontology alignment evaluation initiative 2011, in: Proceedings of the 6th International Ontology matching workshop, Bonn (DE), 2011, pp. 85–110.

[16] J. Euzenat, A. Ferrara, C. Meilicke, A. Nikolov, J. Pane, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, C. Trojahn dos Santos, Results of the ontology alignment evaluation initiative 2010, in: Proceedings of the 5th International Ontology matching workshop, Shanghai (CN), 2010, pp. 85–117. URL: http://oaei.ontologymatching.org/2010/results/oaei2010.pdf.

[17] J. Euzenat, A. Ferrara, L. Hollink, A. Isaac, C. Joslyn, V. Malaisé, C. Meilicke, A. Nikolov, J. Pane, M. Sabou, F. Scharffe, P. Shvaiko, V. Spiliopoulos, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, C. Trojahn dos Santos, G. Vouros, S. Wang, Results of the ontology

alignment evaluation initiative 2009, in: Proceedings of the 4th International Ontology matching workshop, Chantilly (VA, US), 2009, pp. 73–126.

[18] C. Caracciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilicke, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, Results of the ontology alignment evaluation initiative 2008, in: Proceedings of the 3rd Ontology matching workshop, Karlsruhe (DE), 2008, pp. 73–120.

[19] J. Euzenat, A. Isaac, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, W. van Hage, M. Yatskevich, Results of the ontology alignment evaluation initiative 2007, in: Proceedings 2nd International Ontology matching workshop, Busan (KR), 2007, pp. 96–132. URL: http://ceur-ws.org/Vol-304/paper9.pdf.

[20] J. Euzenat, M. Mochol, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, W. R. van Hage, M. Yatskevich, Results of the ontology alignment evaluation initiative 2006, in: Proceedings of the 1st International Ontology matching workshop, Athens (GA, US), 2006, pp. 73–95. URL: http://ceur-ws.org/Vol-225/paper7.pdf.

[21] S. Hertling, J. Portisch, H. Paulheim, Melt - matching evaluation toolkit, in: M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, Y. Sure-Vetter (Eds.), Semantic Systems. The Power of AI and Knowledge Graphs, Springer International Publishing, Cham, 2019, pp. 231–245.

[22] E. Jiménez-Ruiz, T. Saveta, O. Zamazal, S. Hertling, M. Röder, I. Fundulaki, A.-C. N. Ngomo, M. A. Sherif, A. Annane, Z. Bellahsene, S. B. Yahia, G. Diallo, D. Faria, M. Kachroudi, A. Khiat, P. Lambrix, H. Li, M. Mackeprang, M. Mohammadi, M. Rybinski, B. S. Balasubramani, C. Trojahn, Introducing the HOBBIT platform into the Ontology Alignment Evaluation Campaign, in: Proceedings of the 13th International Workshop on Ontology Matching, 2018.

[23] Z. Dragisic, V. Ivanova, H. Li, P. Lambrix, Experiences from the anatomy track in the ontology alignment evaluation initiative, Journal of Biomedical Semantics 8 (2017) 56:1–56:28. URL: https://doi.org/10.1186/s13326-017-0166-5. doi:10.1186/s13326-017-0166-5.

[24] O. Zamazal, V. Svátek, The ten-year ontofarm and its fertilization within the onto-sphere, Web Semantics: Science, Services and Agents on the World Wide Web 43 (2017) 46–53.

[25] C. Meilicke, R. García Castro, F. Freitas, W. van Hage, E. Montiel-Ponsoda, R. Ribeiro de Azevedo, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, A. Tamilin, C. Trojahn, S. Wang, MultiFarm: A benchmark for multilingual ontology matching, Journal of web semantics 15 (2012) 62–68. URL: http://www.sciencedirect.com/science/article/pii/S157082681200039X. doi:10.1016/j.websem.2012.04.001.

[26] P. Buche, J. Cufi, S. Dervaux, J. Dibie, L. Ibanescu, A. Oudot, M. Weber, How to manage incompleteness of nutritional food sources?: A solution using foodon as pivot ontology, Int. J. Agric. Environ. Inf. Syst. 12 (2021) 1–26. URL: https://doi.org/10.4018/ijaeis.20211001.oa4. doi:10.4018/ijaeis.20211001.oa4.

[27] H. Paulheim, S. Hertling, D. Ritze, Towards evaluating interactive ontology matching tools, in: Proceedings of the 10th Extended Semantic Web Conference, Montpellier (FR), 2013, pp. 31–45. URL: http://dx.doi.org/10.1007/978-3-642-38288-8_3.

[28] Z. Dragisic, V. Ivanova, P. Lambrix, D. Faria, E. Jiménez-Ruiz, C. Pesquita, User validation in ontology alignment, in: Proceedings of the 15th International Semantic Web Conference,

Kobe (JP), 2016, pp. 200–217. URL: http://dx.doi.org/10.1007/978-3-319-46523-4_13. doi:10.1007/978-3-319-46523-4_13.

[29] H. Li, Z. Dragisic, D. Faria, V. Ivanova, E. Jiménez-Ruiz, P. Lambrix, C. Pesquita, User validation in ontology alignment: functional assessment and impact, The Knowledge Engineering Review 34 (2019) e15. doi:10.1017/S0269888919000080.

[30] V. Ivanova, P. Lambrix, J. Åberg, Requirements for and evaluation of user support for large-scale ontology alignment, in: Proceedings of the European Semantic Web Conference, 2015, pp. 3–20.

[31] Y. He, J. Chen, H. Dong, E. Jiménez-Ruiz, A. Hadian, I. Horrocks, Machine learning-friendly biomedical datasets for equivalence and subsumption ontology matching, in: U. Sattler, A. Hogan, C. M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d'Amato (Eds.), The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings, volume 13489 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 575–591. URL: https://doi.org/10.1007/978-3-031-19433-7_33. doi:10.1007/978-3-031-19433-7_33.

[32] K. A. Shefchek, N. L. Harris, M. A. Gargano, N. Matentzoglu, D. R. Unni, M. H. Brush, D. Keith, T. Conlin, N. A. Vasilevsky, X. A. Zhang, J. P. Balhoff, L. Babb, S. M. Bello, H. Blau, Y. M. Bradford, S. Carbon, L. Carmody, L. E. Chan, V. Cipriani, A. Cuzick, M. G. D. Rocca, N. A. Dunn, S. Essaid, P. Fey, C. A. Grove, J.-P. F. Gourdine, A. Hamosh, M. A. Harris, I. Helbig, M. E. Hoatlin, M. P. Joachimiak, S. Jupp, K. B. Lett, S. E. Lewis, C. McNamara, Z. M. Pendlington, C. Pilgrim, T. Putman, V. Ravanmehr, J. T. Reese, E. R. Riggs, S. M. C. Robb, P. Roncaglia, J. Seager, E. Segerdell, M. N. Similuk, A. L. Storm, C. Thaxon, A. E. Thessen, J. O. B. Jacobsen, J. A. McMurry, T. Groza, S. Köhler, D. Smedley, P. N. Robinson, C. J. Mungall, M. A. Haendel, M. C. Munoz-Torres, D. Osumi-Sutherland, The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species, Nucleic Acids Research (2020).

[33] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research (2004).

[34] N. Karam, C. Müller-Birn, M. Gleisberg, D. Fichtmüller, R. Tolksdorf, A. Güntsch, A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data, Datenbank-Spektrum 16 (2016) 195–205. URL: https://doi.org/10.1007/s13222-016-0231-8. doi:10.1007/s13222-016-0231-8.

[35] F. Klan, E. Faessler, A. Algergawy, B. König-Ries, U. 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.

[36] N. Karam, A. Khiat, A. Algergawy, M. Sattler, C. Weiland, M. Schmidt, Matching biodiversity and ecology ontologies: challenges and evaluation results, Knowl. Eng. Rev. 35 (2020) e9. URL: https://doi.org/10.1017/S0269888920000132. doi:10.1017/S0269888920000132.

[37] F. Michel, O. Gargominy, S. Tercerie, C. Faron-Zucker, A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF, in: A. Algergawy, N. Karam, F. Klan, C. Jonquet (Eds.), 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.
[38] T. Ashino, Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge, Data Science Journal 9 (2010) 54–61.
[39] M. Wu, P. Hagan, B. Cecconi, S. M. Richard, C. Verhey, R. R. M. S. WG, A collection of crosswalks from fifteen research data schemas to schema.org, 2022. URL: https://doi.org/10.15497/RDA00069. doi:10.15497/RDA00069.
[40] M. Wu, S. M. Richard, C. Verhey, L. J. Castro, B. Cecconi, N. Juty, An analysis of crosswalks from research data schemas to schema.org, Data Intelligence (2022) 1–21. URL: https://doi.org/10.1162/dint_a_00186. doi:10.1162/dint_a_00186.
[41] C. Bizer, J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mende, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, Semantic Web (2012) 1–5.
[42] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 697–706.
[43] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, T. M. Mitchell, Toward an architecture for never-ending language learning, in: Twenty-Fourth AAAI Conference on AI, 2010.
[44] O. Fallatah, Z. Zhang, F. Hopfgartner, The impact of imbalanced class distribution on knowledge graphs matching, in: Proceedings of the 17th International Workshop on Ontology Matching (OM 2022). CEUR-WS, 2022.
[45] O. Fallatah, Z. Zhang, F. 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.
[46] P. Krauss, schemaorg-wikidata-map, https://github.com/okfn-brasil/schemaOrg-Wikidata-Map, 2017.
[47] S. Hertling, H. Paulheim, Dbkwik: extracting and integrating knowledge from thousands of wikis, Knowledge and Information Systems (2019).
[48] S. Hertling, H. Paulheim, Dbkwik: A consolidated knowledge graph from thousands of wikis, in: Proceedings of the International Conference on Big Knowledge, 2018.
[49] T. Saveta, E. Daskalaki, G. Flouris, I. Fundulaki, M. Herschel, A.-C. 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, ACM, New York, NY, USA, 2015, pp. 105–106. URL: http://doi.acm.org/10.1145/2740908.2742729. doi:10.1145/2740908.2742729.
[50] T. D. Doudali, I. Konstantinou, N. K. Doudali, Spaten: a Spatio-Temporal and Textual Big Data Generator, in: IEEE Big Data, 2017, pp. 3416–3421.
[51] C. Strobl, Encyclopedia of GIS, Springer, 2008, pp. 240–245.
[52] Y. He, J. Chen, D. Antonyrajah, I. Horrocks, Bertmap: a bert-based ontology alignment system, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 5684–5691.
[53] J. Chen, Y. He, Y. Geng, E. Jimenez-Ruiz, H. Dong, I. Horrocks, Contextual semantic embeddings for ontology subsumption prediction, arXiv preprint arXiv:2202.09791 (2022).
[54] M. Šatra, O. Zamazal, Towards matching of domain ontologies to cross-domain ontology: Evaluation perspective, in: Proceedings of the 19th International Workshop on Ontology

Matching, 2020.
[55] J. Dabrowski, E. V. Munson, 40 years of searching for the best computer system response time, Interacting with Computers 23 (2011) 555–564. URL: http://www.sciencedirect.com/science/article/pii/S0953543811000579. doi:http://dx.doi.org/10.1016/j.intcom.2011.05.008.
[56] Z. Wang, I. F. Cruz, Agreementmakerdeep results for oaei 2021., in: OM@ ISWC, 2021, pp. 124–130.
[57] D. Faria, M. C. Silva, P. Cotovio, P. Eugénio, C. Pesquita, Matcha and matcha-dl results for oaei 2022, 2022.
[58] E. Jiménez-Ruiz, B. Cuenca Grau, LogMap: Logic-based and scalable ontology matching, in: Proceedings of the 10th International Semantic Web Conference, Bonn (DE), 2011, pp. 273–288.
[59] S. Hertling, H. Paulheim, Atbox results for oaei 2021, in: CEUR Workshop Proceedings, volume 3063, RWTH Aachen, 2021, pp. 137–143.
[60] A. Sharma, A. Patel, S. Jain, Lsmatch results for oaei 2021., in: OM@ ISWC, 2021, pp. 178–184.
[61] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, I. Horrocks, Owl2vec*: Embedding of owl ontologies, Machine Learning 110 (2021) 1813–1845.
[62] J. Portisch, S. Hertling, H. Paulheim, Visual analysis of ontology matching results with the melt dashboard, in: European Semantic Web Conference, 2020, pp. 186–190.

Linköping, Jena, Lisboa, Heraklion, Mannheim, Montpellier, Oslo, London, Berlin, Trento, Toulouse, Prague, Manhattan, Dublin, Oxford
December 2022