Results of the Ontology Alignment Evaluation Initiative 2023

Mina Abd Nikooie Pour[1], Alsayed Algergawy[2,3], Patrice Buche[4], Leyla J. Castro[5], Jiaoyan Chen[6], Adrien Coulet[7,8], Julien Cufi[4], Hang Dong[9], Omaima Fallatah[10], Daniel Faria[11], Irini Fundulaki[12], Sven Hertling[13], Yuan He[9], Ian Horrocks[9], Martin Huschka[14], Liliana Ibanescu[15], Sarika Jain[16], Ernesto Jiménez-Ruiz[17], Naouel Karam[18], Patrick Lambrix[1], Huanyu Li[1], Ying Li[1], Pierre Monnin[19], Engy Nasr[20], Heiko Paulheim[13], Catia Pesquita[21], Tzanina Saveta[12], Pavel Shvaiko[22], Guilherme Sousa[23], Cassia Trojahn[23], Jana Vatascinova[24], Mingfang Wu[25], Beyza Yaman[26], Ondrej Zamazal[24] and Lu Zhou[27]

[1]*Linköping University & Swedish e-Science Research Centre, Linköping, Sweden*
[2]*Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Germany*
[3]*Chair of Data and Knowledge Engineering, University of Passau, Germany*
[4]*UMR IATE, INRAE, University of Montpellier, France*
[5]*ZB MED Information Centre for Life Sciences, Germany*
[6]*Department of Computer Science, The University of Manchester, UK*
[7]*Inria Paris, France*
[8]*Centre de Recherche des Cordeliers, Inserm, Université Paris Cité, Sorbonne Université, France*
[9]*Department of Computer Science, University of Oxford, UK*
[10]*Department of Data Science, Umm Al-Qura University, Saudi Arabia*
[11]*INESC-ID / IST, University of Lisbon, Portugal*
[12]*Institute of Computer Science-FORTH, Heraklion, Greece*
[13]*Data and Web Science Group, University of Mannheim, Germany*
[14]*Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institut, EMI, Germany*
[15]*Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA Paris-Saclay, France*
[16]*National Institute of Technology Kurukshetra, India*
[17]*City, University of London, UK & SIRIUS, University of Oslo, Norway*
[18]*Institute for Applied Informatics, University of Leipzig, Germany*
[19]*University Côte d'Azur, Inria, CNRS, I3S, France*
[20]*Albert Ludwig University of Freiburg, Germany*
[21]*LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[22]*Trentino Digitale SpA, Trento, Italy*
[23]*Institut de Recherche en Informatique de Toulouse, France*
[24]*Prague University of Economics and Business, Czech Republic*
[25]*Australian Research Data Commons*
[26]*ADAPT Centre, Trinity College Dublin*
[27]*Flatfee Corp, 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 2023 campaign offered 15 tracks and was attended by 16 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 to allow anyone to conclude 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, co-located with ISWC [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21], which this year took place in Athens, Greece[2].

Since 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 their preferred platform. Since last year, the MELT framework [22] has been adopted to facilitate the SEALS and HOBBIT wrapping and evaluation. This year, most tracks have adopted MELT as their evaluation platform.

This paper synthesizes the 2023 evaluation campaign and introduces the results provided in the participants' papers. 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 2023: The 18th International Workshop on Ontology Matching collocated with the 22nd International Semantic Web Conference ISWC-2023 November 7th, 2023, Athens, Greece

CEUR Workshop Proceedings (CEUR-WS.org)

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

[2] http://om2023.ontologymatching.org

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

2. Methodology

2.1. Evaluation platforms

The OAEI evaluation was conducted 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 of 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 [23].

The **MELT framework**[5] [22] 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). Starting with this year, the MELT framework also supports the SSSOM [24] format. Therefore, systems producing an alignment in the SSSOM format can be evaluated as well.

All platforms compute the standard evaluation metrics against the reference alignments: precision, recall, and F-measure. In test cases requiring different evaluation modalities, 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 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 2023, 12 systems were submitted as MELT Web docker container, 3 systems were submitted as SEALS package, 1 system was uploaded to the HOBBIT platform, and one system implemented the Web interface directly and provided hosting for the system.

In this year we also allowed to submit alignment files in addition to the executable system in case it requires substantial hardware or software resources.

[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 2023 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}, 2023. 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^{st}. The execution phase was terminated on September 30^{th}, 2023, 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 15 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.

Table 1
Tracks in OAEI 2023.

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		√	
pharmacogenomics	OWL	=, <, >, Close, Related	[0 1]	open	EN			√

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 [25].

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 4.2.

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

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

domain of organizing conferences. The dataset and its usage are described in [26]. This year we 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 [27] 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. This year, we report the results on different ontologies (i).

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. The evalua-

tion 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.

3.4. 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*.

This year the track run with two data sets from the conference domain: Conference and Populated Conference, as the other complex sub-tracks (**Hydrography**, **GeoLink**, **Populated GeoLink Populated Enslaved**, and **Taxon** datasets) have been discontinued.

The **Conference** dataset comprises three ontologies: cmt, conference, and ekaw from the conference dataset. The reference alignment was created as a consensus between experts. To allow matchers which rely on instances to participate over the Conference complex track, the **Populated Conference** data set is composed of 5 conference ontologies populated with more or less common instances, resulting in 6 datasets: (6 versions on the repository: v0, v20, v40, v60, v80 and v100). Details on the population and evaluation modalities are available[10].

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.

3.5. Food

The Food Nutritional Composition track aims at finding alignments between food concepts from CIQUAL[11], the French food nutritional composition database, and food concepts from SIREN[12], the Scientific Information and Retrieval Exchange Network of the US Food and Drug administration. Foods from both databases are described in LanguaL[13], a well-known multilingual thesaurus using faceted classification. LanguaL stands for "Langua aLimentaria" or "language of food"; more than 40,000 foods used in food composition databases are described using LanguaL.

In [28], a method to provide OWL modelling of food concepts from both datasets, CIQUAL[14] and SIREN [15], and a gold standard are presented.

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 [29, 30, 31]. The evaluation thus focuses on how interaction

[10] https://framagit.org/IRIT_UT2J/conference-dataset-population
[11] https://ciqual.anses.fr/
[12] http://langual.org/langual_indexed_datasets.asp
[13] https://www.langual.org/default.asp
[14] https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.15454/6CEYU3
[15] https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.15454/5LLGVY

with the user improves the matching results. Currently, this track does not evaluate the user experience or the user interfaces of the systems [32, 30].

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, telling 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 [33] incorporates both *equivalence* and *subsumption* ontology matching (OM) tasks for biomedical ontologies, with ground truth (equivalence) mappings extracted from Mondo [34] and UMLS [35] (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 [33], 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.e., *unsupervised* setting with no training mappings, and *semi-supervised* setting with 30% ground truth mappings for training/validation. In the 2023 edition, we made several significant updates:

- *Logical module enrichment*: we adopted locality-based logical modules [36] to enrich the existing pruned ontologies to provide more contexts for alignment; the added entities are annotated as *"not used in alignment"* and will be ignored in evaluation.

- *Bio-LLM sub-track*: To support more efficient evaluation of large language model-based OM, we introduced a special sub-track called Bio-LLM [37], which consists of challenging

subsets of NCIT-DOID and SNOMED-FMA (Body) datasets, along with tailored evaluation metrics.

– *Simplified file structure and task settings*: We also re-organised the structure of the dataset files, and reduced the task settings such that the unsupervised and semi-supervised settings share the same testing set for ranking evaluation.

For evaluation, in [33] 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 to distinguish 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. For the special sub-track Bio-LLM introduced in the 2023 edition, both matching and ranking metrics are used but they are tailored to the subsets, along with an additional metric called rejection rate to examine if systems can reject all plausible mappings for entities that actually have no alignment.

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[16]	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`[17]	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

We adopted a flexible way of 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 are 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. In the final result tables, we used superscripts †, ‡, and * to indicate

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

that the results came from MELT, source code implementation, and direct result submission, respectively. All our evaluations were conducted with the DeepOnto[18] [38] 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[19] (The German Federation for Biological Data) alongside its successor NFDI4Biodiversity[20] and the AquaDiva[21] projects, which aim at providing semantically enriched data management solutions for data capture, annotation, indexing and search [39, 40, 41]. In this track, we aim to motivate ontology matching systems to work on matching ontologies and thesauri used in the biodiversity and ecology domains, available via the BiodivPortal ontology repository[22]. For the current edition, we kept the matching task between the Environment Ontology (ENVO) and the Semantic Web for Earth and Environment Technology Ontology (SWEET) as these two ontologies have frequent updates.

In 2021, we added a task to align two biological taxonomies with rather different but complementary scopes: the well-known NCBI taxonomy (NCBITAXON), and TAXREF-LD [42]. No matching system was able to achieve this matching task due to the large size of the considered taxonomies. To cope with this issue since last year edition, we split the large matching task into a set of smaller, more manageable subtasks through the use of modularization [43]. We obtained six groups corresponding to the kingdoms: Animalia, Bacteria, Chromista, Fungi, Plantae and Protozoa, leading to six well balanced matching subtasks. In 2023, we partnered with the EcoPortal project[23] to include two new matching tasks involving important thesauri in environmental sciences (originally developed in SKOS): finding alignments between the Macroalgae Traits Thesaurus (MACROALGAE) and the Macrozoobenthos Traits Thesaurus (MACROZOOBENTHOS) and between the Fish Traits Thesaurus (FISH) and the Zooplankton Traits Thesaurus (ZOOPLANKTON). Table 4 presents detailed information about the ontologies and thesauri used in this year's edition.

3.9. Material Sciences and Engineering (MSE)

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

[18] https://krr-oxford.github.io/DeepOnto/#/
[19] www.gfbio.org
[20] www.nfdi4biodiversity.org/en/
[21] www.aquadiva.uni-jena.de
[22] biodivportal.gfbio.org/
[23] ecoportal.lifewatch.eu/

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	-
MACROALGAE	SKOS	2018-10-02	-	109
MACROZOOBENTHOS	SKOS	2023-07-11 (v1.1)	-	128
FISH	SKOS	2015-03-11	-	146
ZOOPLANKTON	SKOS	2019-05-27	-	57
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	-

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		

principles with only partial overlap with each other. The current version v1.1[24] of the MSE track includes three test cases summarized 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.

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[44] 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 more efficient creation of the manual reference alignment in the First Test Case, see Table 5.

[24] https://github.com/EngyNasr/MSE-Benchmark/releases/tag/v1.1

The MatOnto Ontology v2.1[25] (847 classes, 96 properties and 131 individuals) bases on the upper-level ontology bfo2[26]. The Elementary Multiperspective Material Ontology (EMMO v1.0.0-alpha2)[27], is a standard representational ontology framework based on current materials modeling and characterization knowledge incorporating an upper-, mid- and domain-level (451 classes, 35 properties). For every test case, a manual reference alignment $\boldsymbol{R}$ was created in close cooperation with MSE domain experts.

The evaluation was performed using the MELT platform on a Windows 11 system with Intel Core i7-1260P CPU @2.10GHz and 16 GB RAM. Every participating system was executed in its standard setting, and we compared precision, recall, and F-measure as well as the computation time. No background knowledge was used for evaluation.

3.10. Crosswalks Data Schema Matching

This track was introduced in 2022, aiming 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 schemes to Schema.org [45, 46]. It is based on the work carried out by the Research Data Alliance (RDA) Research Metadata Schemas Working Group. The collection serves as a reference for data repositories when they develop their crosswalks, as well as an indication of semantic interoperability among the schemas.

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

Using as a reference the manually established correspondences, the evaluation here will be based on the well-known measures of precision, recall, and F-measure. 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.

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 [47], YAGO [48] and NELL [49]. The dataset used for the evaluation is generated from DBpedia and

[25] https://raw.githubusercontent.com/iNovexIrad/MatOnto-Ontologies/master/matonto-release.ttl
[26] http://purl.obolibrary.org/obo/bfo/2.0/bfo.owl
[27] 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 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 [50]. The numbers of entities in the four KG datasets are illustrated in Table 6.

The NELL and DBpedia benchmark [51] 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 [52] and expanded according to OAEI standard as part of [50].

With 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 a 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[28] in the course of the DBkWik project [53, 54]. 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.

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 interwiki 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

[28] https://www.wikia.com/

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 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[29].

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[30].

3.13. SPIMBENCH and Link Discovery

This year, only LogMap has participated in the SPIMBENCH and Link Discovery tracks. The organizers then made the decision not to run these tracks this year.

3.14. Pharmacogenomics

The Pharmacogenomics track is a new track proposed for OAEI 2023 that focuses on matching knowledge units from the pharmacogenomics domain. These units are n-ary tuples – so-called "pharmacogenomic relationships" – and involve drugs, genetic factors, and phenotypes (see Figure 1). A pharmacogenomic tuple states that patients being treated by the specified drugs while having the specified genetic factors may experience the given phenotypes.

In the Semantic Web formalisms, only binary predicates exist. That is why pharmacogenomic tuples are reified: tuples become individuals that are linked to their components with binary

[29] https://github.com/dwslab/melt/tree/master/examples/kgEvalCli
[30] http://oaei.ontologymatching.org/2019/results/knowledgegraph/kgBaselineMatchers.zip

predicates (Figure 1(c)). Hence, the task of matching pharmacogenomic tuples is [55]:

- An *instance matching task* that aims at finding alignments between individuals representing reified tuples;
- A *structure-based matching task* in which neighbors of reified tuples are compared to conclude the potential alignment between tuples. Recall that no other information exist about the tuples except their neighbors (*e.g.*, no labels).

To illustrate, two tuples involving the same drugs, genetic factors, and phenotypes will thus be connected to the same neighbors and are expected to be detected as identical.

Beside the arity of tuples, matchers need to face their potential incompleteness (*e.g.*, missing drugs) and heterogeneity (*e.g.*, a gene version like `CYP2C9*4` is more specific than the gene itself `CYP2C9`, the phenotype `hemorrhage` is more specific than the phenotype `vascular disorders`). *Different types of alignments* are thus expected to be identified between pharmacogenomic tuples, which is somehow unusual in an instance matching task. The Pharmacogenomics track features the identification of identical tuples ($=$), equivalent tuples (`Close`), tuples being more specific ($<$) or more general ($>$) than others, and tuples being related to some extent (`Related`). See [55, 56] for a detailed definition of these different alignment types between individuals.

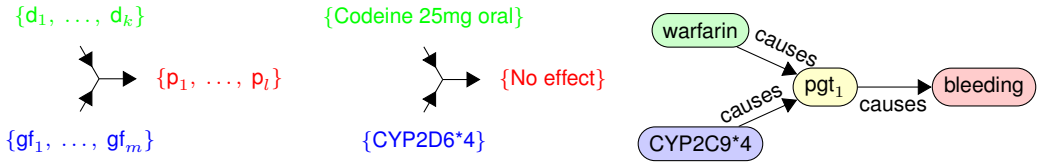
To perform this alignment task, matchers can rely on additional background knowledge about components of pharmacogenomic tuples. This knowledge includes ontology classes instanciated by components of tuples (*i.e.* drugs, genetic factors, phenotypes) and their hierarchical organization, `partOf` links between gene versions and genes, `sameAs` links between identical drugs, genes, or phenotypes, and `dependsOn` links between complex phenotypes and their components (*e.g.*, "warfarin-induced bleeding" depends on "warfarin" and on "bleeding").

To evaluate matchers and their scalability, the Pharmacogenomics track comprises three tasks involving respectively 10, 50, and 100% of the 50,435 pharmacogenomic tuples represented within the PGxLOD knowledge graph[31] [57]. For each task, the selected pharmacogenomic tuples are evenly split into two ontologies to match. To take into account the specificity of the different alignment types that are expected, matchers are evaluated through two settings:

Fine-grained setting Only alignments of the exact type expected in the reference are considered correct. To illustrate, an output alignment $(e_1, =, e_2)$ where $(e_1, \texttt{Close}, e_2)$ was expected will be considered as incorrect. Precision, Recall, and F1-score are computed for each type of alignment.

Coarse-grained setting Any type of alignment between entities expected to be aligned will be considered as correct. To illustrate, an output alignment $(e_1, =, e_2)$ where $(e_1, \texttt{Close}, e_2)$ was expected will be considered as correct. Precision, Recall, and F1-score are computed globally accordingly.

[31] https://pgxlod.loria.fr/



(a) Abstract relationship (b) Example relationship (c) Reified relationship

Figure 1: Graphical representation of an abstract (1(a)), an example (1(b)), and a reified (1(c)) pharmacogenomic relationships. The example relationship states that patients having the "*4" version of the $CYP2D6$ gene will not experience the expected effect of codeine. gf stands for genetic factor, d for drug and p for phenotype.

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 16 participating systems. Table 8 lists the participants and the tracks in which they competed. It is worth mentioning that the Bio-ML track has additional participants (e.g., BERTMap [58] and BERTSubs [59]) 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 summarize the results for each track.

4.2. Anatomy

The results for the Anatomy track are shown in Table 9. Of the 9 systems participating in the Anatomy track, 8 achieved an F-measure higher than the StringEquiv baseline. Two systems were first-time participants (SORBETMatcher, OLaLa). 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 exception were LogMapBio which increase in F-measure (from 0.895 to 0.898) and Matcha increased in recall+ (from 0.817 to 0.818), in size (from 1482 to 1484). In terms of run time, 5 out of 9 systems computed an alignment in less than 100 seconds. LogMapLt remains the system with the shortest runtime. Regarding quality, Matcha achieved the highest F-measure (0.941) and recall+ (0.818), but four other systems obtained an F-measure above 0.88 (OLaLa, SORBETMatcher, LogMapBio, and LogMap) which is at least as good as the best systems in OAEI 2007-2010. Like in previous years, there is no significant correlation between the quality of the generated alignment and the run time. Two systems produced coherent alignments (LogMapBio and LogMap).

Table 8
Participants and the status of their submissions.

System	ALIN	AMD	GraphMatcher	LogMap	LogMapBio	LogMapLt	LogMapKG	LSMatch	LSMatch-Multilingual	Matcha	MatchaC	Matcha-DL	OLaLa	PropMatch	SORBETMatcher	TOMATO	Total=16
anatomy	●	●	○	●	●	●	○	●	○	●	○	○	●	○	●	○	9
conference	●	●	●	●	○	●	○	●	○	●	○	○	●	●	●	●	11
multifarm	○	○	○	●	○	●	○	○	●	●	○	○	○	○	○	○	4
complex	○	○	○	●	○	◐	○	○	○	○	◐	○	○	○	○	○	3
food	○	○	○	●	○	●	○	○	○	●	○	○	●	○	○	○	4
interactive	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	2
bio-ML	○	◐	○	◐	◐	◐	○	○	○	○	○	◐	◐	○	◐	○	7
biodiv	○	○	○	●	○	●	●	○	○	◐	○	○	●	○	○	○	5
mse	○	◐	○	●	○	●	○	○	○	●	○	○	○	○	○	○	4
common knowl. graph	○	●	○	●	○	●	●	●	○	●	○	○	●	○	○	○	7
crosswalks	○	○	○	●	●	●	●	○	○	●	○	○	○	○	○	○	5
knowledge graph	○	○	○	●	○	●	○	●	○	●	○	○	●	○	●	○	6
spimbench	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	1
link discovery	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	1
pharmacogenomics	○	○	○	◐	◐	◐	◐	○	○	◐	○	○	○	○	○	○	5
total	3	5	1	15	4	12	4	4	1	10	1	1	7	1	4	1	

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	54	1484	0.951	0.941	0.931	0.818	-
OLaLa	9259	1470	0.924	0.91	0.896	0.726	-
SORBETMatcher	4032	1470	0.923	0.909	0.895	0.724	-
LogMapBio	1945	1578	0.88	0.898	0.916	0.778	√
LogMap	10	1402	0.917	0.881	0.848	0.602	√
AMD	36	1282	0.938	0.86	0.794	0.461	-
ALIN	374	1159	0.984	0.852	0.752	0.501	-
LogMapLt	3	1147	0.962	0.828	0.728	0.288	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
LSMatch	16	1009	0.952	0.761	0.634	0.037	-

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 the conference track's web page.

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 the number of incoherent alignments. Conser.V. means the total number of all conservative principle violations. Consist.V. means the 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.
GraphMatcher	0.71	0.72	0.74	0.76	0.77	8	172	85
SORBETMatcher	0.73	0.7	0.66	0.63	0.61	7	43	73
LogMap	0.76	0.71	0.64	0.59	0.56	0	2	0
Matcha	0.62	0.62	0.62	0.62	0.62	7	90	81
OLaLa	0.59	0.59	0.6	0.61	0.61	11	199	184
ALIN	0.82	0.7	0.57	0.48	0.44	0	2	0
edna	0.74	0.66	0.56	0.49	0.45			
LogMapLt	0.68	0.62	0.56	0.5	0.47	0	21	0
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	3	97	18
StringEquiv	0.76	0.65	0.53	0.45	0.41			
TOMATO	0.57	0.55	0.52	0.49	0.47	13	361	203
PropMatch	0.86	0.29	0.15	0.1	0.08	0	0	0

With regard to two baselines we can group tools according to system's position: six systems outperformed above both baselines (GraphMatcher, SORBETMatcher, LogMap, Matcha, ALIN, and OLaLa); three systems performed better than StringEquiv baseline (LogMapLt, AMD, and LSMatch), and two systems performed worse than both baselines (TOMATO, and PropMatch). Four matchers (AMD, ALIN, LSMatch, and SORBETMatcher) do not match properties at all. On the other side, PropMatch does not match classes at all, while it dominates in matching properties. Naturally, this has a negative effect on their overall performance.

Several systems use reference alignments to a certain extent (GraphMatcher uses it in its 5-fold cross-validation and TOMATO uses it in its sampling process) for their model learning, as has happened with some ML-based systems in the past. These systems describe their usage in their system papers.

The performance of all matching systems regarding their precision, recall and F_1-measure is plotted in Figure 2. 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 11 alignment systems, 6 (ALIN, AMD, LogMapLt, LSMatch, SORBETMatcher, and TOMATO) use 1.0 as the confidence value for all matches they identify. The remaining 5 systems (GraphMatcher, LogMap, Matcha, OLaLa, and PropMatch) have a wide variation of confidence values.

When comparing the performance of the systems on uncertain reference alignments versus the sharp versions, it is evident that in the discrete cases, all systems either performed at the same

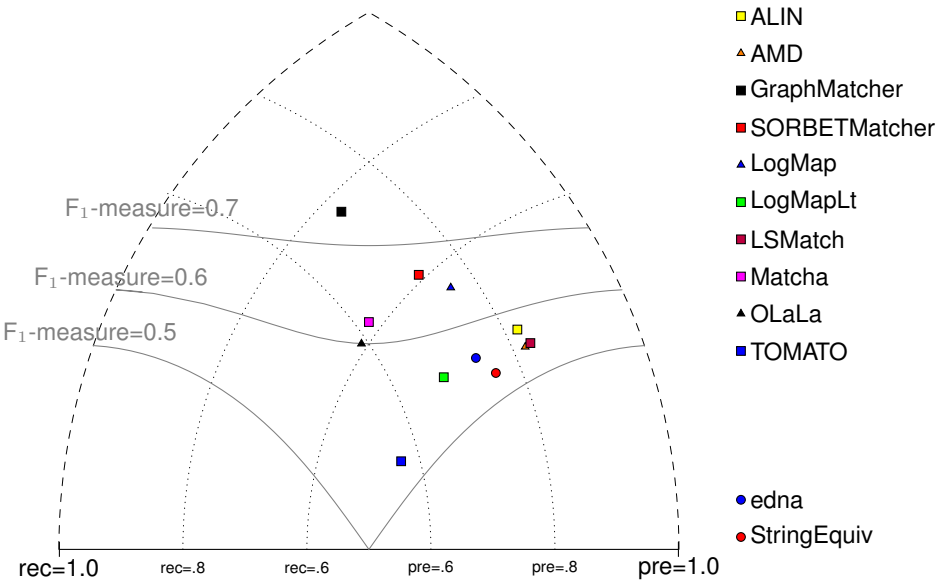


Figure 2: Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F_1-measure are depicted by areas bordered by corresponding lines F_1-measure=0.[5|6|7].

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

System	Sharp			Discrete			Continuous		
	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
AMD	0.87	0.58	0.43	0.87	0.66	0.53	0.86	0.67	0.55
GraphMatcher	0.76	0.78	0.80	0.64	0.73	0.84	0.65	0.73	0.84
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	0.88	0.66	0.53	0.88	0.67	0.54
Matcha	0.70	0.67	0.64	0.62	0.68	0.76	0.64	0.69	0.74
OLaLa	0.56	0.61	0.66	0.58	0.65	0.75	0.57	0.63	0.70
PropMatch	0.83	0.15	0.08	0.76	0.16	0.09	0.73	0.13	0.07
SORBETMatcher	0.78	0.70	0.64	0.71	0.71	0.72	0.69	0.70	0.72
TOMATO	0.61	0.55	0.61	0.61	0.62	0.63	0.60	0.62	0.64

level as before or showed improvements in terms of F-measure. The changes in F-measure for discrete cases ranged from 1 to 16 percent above those in the sharp reference alignment. Notably, LSMatch exhibited the most significant performance surge at (16%), closely followed by ALIN at (15%) and AMD at (14%). This substantial improvement was primarily driven by increased recall, a result of having fewer 'controversial' matches in the uncertain version of the reference alignment.

In contrast, systems with confidence values consistently set at 1.0 delivered very similar performance, regardless of whether a discrete or continuous evaluation methodology was applied. This was due to their proficiency in identifying matches with which experts had a high degree of agreement, while the matches they missed were typically the more contentious ones. Notably, GraphMatcher stood out by producing the highest F-measure under both continuous (73%) and discrete (73%) evaluation methodologies. This indicates that the system's confidence evaluation effectively reflects the consensus among experts in this task. However, it's worth noting that GraphMatcher experienced relatively small drops in F-measure when transitioning from discrete to continuous evaluation, primarily due to a decrease in precision.

In addition to the above findings, eight systems that participated this year were also part of the previous year's evaluation, allowing for some valuable comparisons over time. Among these, six systems demonstrated remarkable stability in their F-measures when assessed against uncertain reference alignments. However, two systems, Matcha and TOMATO, exhibited significant improvements this year. Matcha's F-measure jumped from (12%) in continuous and (14%) in discrete last year to (63%) in continuous and (65%) in discrete this year, primarily due to an increase in precision. Similarly, TOMATO saw a substantial increase in F-measure from (15%) last year to (62%) this year, both in continuous and discrete evaluations. These improvements mark significant progress for these two systems compared to the previous year.

OLaLa, PropMatch, and SORBETMatcher are three new systems participating this year. OLaLa has shown notable performance improvements, with a (4%) increase in the discrete case and a (2%) increase in the continuous case concerning F-measure when compared to the sharp reference alignment. OLaLa's F-measure has risen from (61%) to (65%) in the discrete case and to (63%) in the continuous case. This improvement is primarily attributed to an increase in recall.

On the other hand, PropMatch and SORBETMatcher have demonstrated similar performance in both discrete and continuous cases when compared to the sharp reference alignment in terms of F-measure. Notably, PropMatch exhibits consistently lower precision and recall across the three different versions of the reference alignment, primarily due to its narrow focus on property matching.

4.4. Multifarm

This year, 4 systems have registered to participate in the Multifarm track:LSMatch-Multilingual, LogMap, LogMapLt, and Matcha. The number of participating tools is stable with respect to the last 4 campaigns (5 in 2022, 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 CIDER-LM. 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 12, 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).

The systems have been executed on a Ubuntu Linux machine configured with 32GB of RAM

Table 12
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.
Matcha	2	.37	.08	.04
LSMatch-Multilingual	46	.68	.47	.36
LogMap	9	.72	.44	.31
LogMapLt	227	.24	.038	.02

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 to complete the 55 x 24 matching tasks: LSMatch-Multilingual (46 min), LogMap (9 minutes), LogMapLt (227 minutes), and Matcha(2 minutes). When we compare the times to last year's campaign, we can see that LogMap has a stable 9 min execution, whereas LogMapLt saw a decline in timing from 175 min to 227 min, and LSMatch-Multilingual improved the timing from 69 min to 46 min. Since the other tools are participating for the first time, their timings are incomparable. These measurements only indicate the time the systems require to finish the task in a common environment. LSMatch-Multilingual outperforms all other systems in terms of F-measure (0.47) and recall (0.36), and LogMap outperforms all other systems in terms of Precision (0.72).

It is seen that a similar number of systems have participated in the campaign through the years. However, there is a dynamicity of the tools, such that, each year, participating tools vary. In 2023, we had four systems participating in the campaign where 1 of them was a new system, and 3 were long-term participating systems. As observed in several campaigns, still, all systems still privilege precision in detriment to recall (recall below 0.50), and the results are below the ones obtained for the Conference's original dataset.

4.5. Complex Matching

Unfortunately, this track has not attracted many participants in the last two years. This year, MatchaC, LogMap and LogMapLt have been registered to participate. While LogMapLt and LogMap are dedicated to generating simple correspondences, only LogMap was able to generate non-empty alignments. MatchaC, the only system specifically designed to generate expressive correspondences, had some problems dealing with the datasets and was not able to generate any valid alignment.

4.6. Food

This is the second year of the track and five systems were registered: AMD, LogMap, LogMapLt, OLaLa and Matcha. AMD encountered an error during launch, so evaluation results only available for four of the matchers are presented in Table 13.

The test case food v2 evaluates matching systems regarding their capability to find "equal" (=) and "subclass" relation ($<$) correspondences between the CIQUAL ontology and the SIREN ontology. All evaluated systems compute the alignment in less than a minute except OLaLa.

Table 13
Food track results per matcher. Time is measured in seconds.

System	Corresp.	Precision	Recall	F1-measure	Time(s)
"equal" (=) relation					
OLaLa	1185	0.0337	0.5479	0.06359	73920
LogMap	15	0.1333	0.0274	0.0454	14
LogMapLt	15	0.1333	0.0274	0.0454	5
Matcha	335	0.0656	0.3013	0.1078	53
"subclass" relation (<) relation					
LogMap	15	0	0	0	15
LogMapLt	15	0	0	0	6
Matcha	335	0	0	0	53

LogMapLt stands out for its very fast calculation time of 5s (resp. 6s) to find "equal" (resp. "subclass" relation correspondences). Concerning "equal" (=) relation correspondences, LogMap and LogMapLt have better precision than Matcha and OLaLa. However, LogMap's recall is 20 (resp. 11 times) less than OLaLa's (resp. Matcha's) one. Matcha is the best-performing participant in the FNC test case in terms of precision and F1-measure. None of the matching systems are able to find "subclass" relation (<) correspondences.

4.7. Interactive matching

This year, two systems (ALIN, and LogMap) participated in the Interactive matching track. Their results are shown in Table 14 and Figure 3 for both the Anatomy and Conference datasets.

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

- The performance of the system: Precision (Prec.), Recall (Rec.), and F-measure (F-m.) with respect to the fixed reference alignment, as well as Recall+ (Rec.+) for the Anatomy task. To facilitate the assessment of the impact of user interactions, we also provide the performance results from the original tracks, without interaction (line with Error NI).

- To ascertain the impact of the oracle errors, we provide the performance of the system with respect to the oracle (i.e., the reference alignment as modified by the errors introduced by the oracle: Precision oracle (Prec. oracle), Recall oracle (Rec. oracle) and F-measure oracle (F-m. oracle). For a perfect oracle, these values match the actual performance of the system.

- Total requests (Tot Reqs.) represents the number of distinct user interactions with the tool, where each interaction can contain one to three conflicting correspondences, that could be analyzed 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).

Table 14
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	514	1453	1.0	1.0
	0.1	0.91	0.9	0.904	0.759	0.987	0.922	0.953	491	1390	0.666	0.975
	0.2	0.846	0.883	0.864	0.73	0.988	0.924	0.955	476	1352	0.469	0.947
	0.3	0.792	0.864	0.826	0.699	0.988	0.925	0.955	466	1329	0.342	0.912
LogMap	NI	0.915	0.848	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	0.966	0.831	0.894	0.569	0.971	0.803	0.879	388	1164	0.75	0.967
	0.2	0.952	0.822	0.882	0.549	0.951	0.764	0.847	388	1164	0.573	0.925
	0.3	0.936	0.819	0.873	0.545	0.925	0.723	0.812	388	1164	0.43	0.876
Conference Dataset												
ALIN	NI	0.874	0.456	0.599	–	–	–	–	–	–	–	–
	0.0	0.919	0.744	0.822	–	0.919	0.744	0.822	274	815	1.0	1.0
	0.1	0.705	0.7	0.702	–	0.934	0.773	0.846	264	785	0.504	0.988
	0.2	0.569	0.663	0.612	–	0.944	0.796	0.863	291	764	0.305	0.977
	0.3	0.462	0.634	0.535	–	0.952	0.816	0.879	250	743	0.196	0.955
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	0.851	0.599	0.703	–	0.862	0.577	0.691	82	246	0.703	0.983
	0.2	0.824	0.592	0.689	–	0.841	0.551	0.666	82	246	0.526	0.952
	0.3	0.793	0.581	0.671	–	0.804	0.519	0.631	82	246	0.36	0.897

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

– 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 of its high

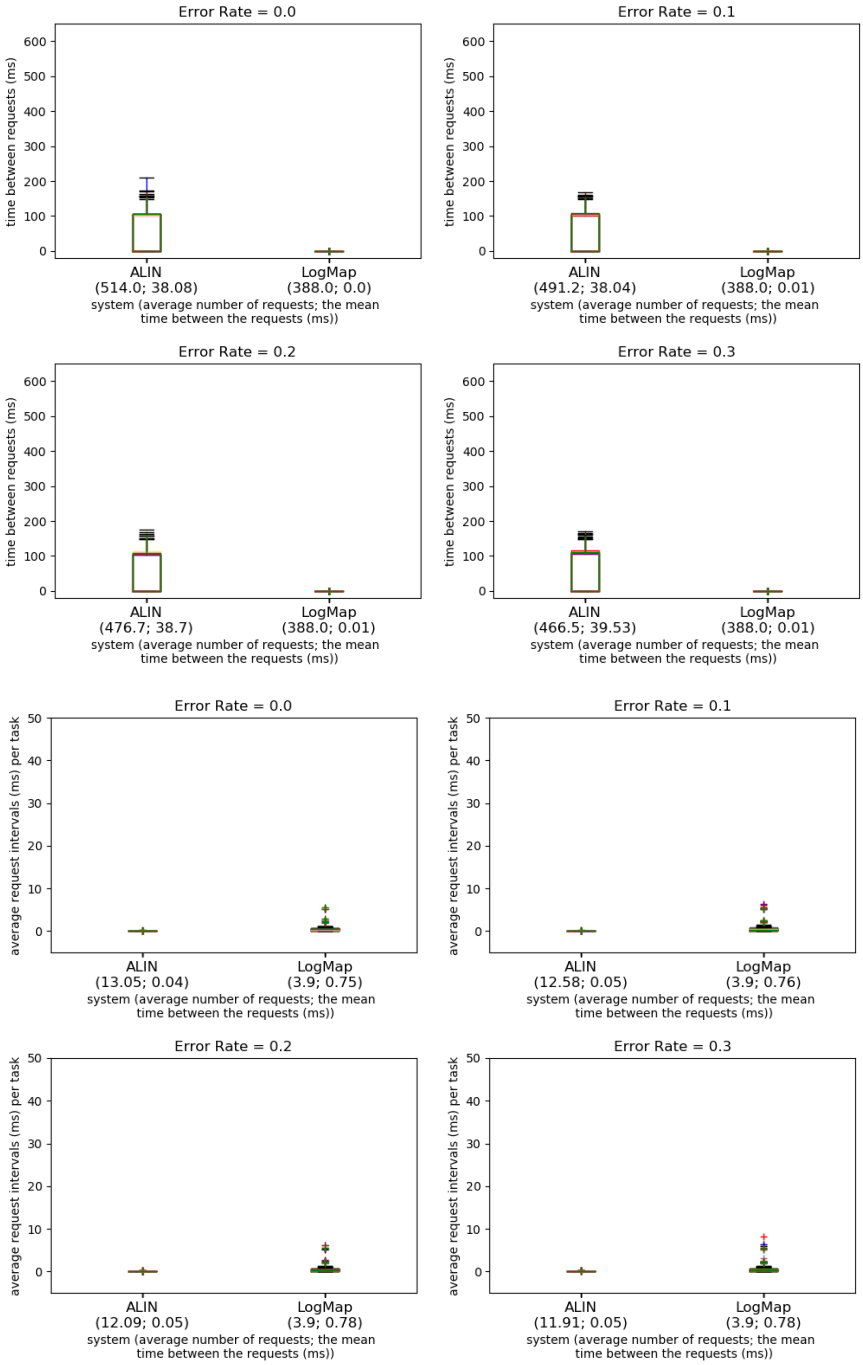


Figure 3: Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

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 [60]: Shneiderman and Seow take different approaches to categorize the response times taking a task-centered view and a user-centered view respectively. According to task complexity, Shneiderman defines response time in four categories: typing, mouse movement (50-150 ms), simple frequent tasks (1 s), common tasks (2-4 s) and complex tasks (8-12 s). While Seow's definition of response time is based on the user expectations towards the execution of a task: instantaneous (100-200 ms), immediate (0.5-1 s), continuous (2-5 s), captive (7-10 s). Ontology alignment is a cognitively demanding task and can fall into the third or fourth categories in both models. In this regard the response times (request intervals as we call them above) observed in all datasets fall into the tolerable and acceptable response times, and even into the first categories, in both models. The request intervals for 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 five tables for equivalence matching, five tables for subsumption matching, and two tables for Bio-LLM, where each table corresponds to an OM pair and includes results of both the unsupervised and semi-supervised settings. For the full results, please refer to the OAEI 2023 Bio-ML website[32].

Briefly, we have the following participants for equivalence matching: *(i)* machine learning-based systems including BERTMap, BERTMapLt [58], AMD [61], Matcha, Matcha-DL [62], OLala, and SORBETMatcher [63]; and *(ii)* traditional systems including LogMap, LogMapBio, and LogMapLt [36]. For equivalence matching, top performing systems vary across different tasks, with LogMapBio attaining the best F1 on 2 out of 5 unsupervised tasks, and AMD, BERTMap, and SORBETMatcher attaining the best F1 on each of the remaining three, respectively. BERTMap also attains the best ranking scores of all tasks, though most systems do not provide ranking results in equivalence matching. For subsumption matching, all the participating systems are machine learning-based, including BERTSubs (IC) [59], OWL2Vec*+RF [64], SORBETMatcher [63], and Word2Vec+Random Forest (RF). SORBETMatcher attains the best MRR on 3 out of 5 subsumption tasks, while BERTSubs (IC) and OWL2Vec*+RF attain the best MRR on each of

[32] https://www.cs.ox.ac.uk/isg/projects/ConCur/oaei/2023/, permanent link at Internet Wayback Machine, https://web.archive.org/web/20231120093331/https://www.cs.ox.ac.uk/isg/projects/ConCur/oaei/2023/

the remaining two, respectively. Overall, the 2023 edition attracted more machine learning-based participants, which matches the original purpose of Bio-ML, while LogMap variants are the only symbolic participants.

4.9. Biodiversity and Ecology

This year, five matching systems (LogMap, LogMapLt, LogMapKG, Matcha and OLaLa) managed to generate an output for all of the track tasks, except Matcha failed to achieve alignment for the envo-sweet task. 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 15.

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, OLaLa took the longer. Regarding the ENVO-SWEET task, only OLaLa and the LogMap family systems achieved it with a similar performance to last year. The MACROALGAE-MACROZOOBENTHOS and FISH-ZOOPLANKTON matching tasks involve resources developed in SKOS. 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. The systems that did not perform well in this task did map a large number of dissimilar concepts that happen to have similar URIs. All systems performed well on most NCBITAXON-TAXREF-LD subtasks, with slightly the same levels of precision and recall. Overall, in this year's evaluation, the number of participating systems decreased and the performance of the successful ones remained similar.

4.10. Material Sciences and Engineering (MSE)

This year four systems registered on the MSE track, each of which was used for evaluation with the three test cases of the MSE benchmark. AMD produced errors and an empty alignment file, so results are only available for three of the matchers: LogMap, LogMapLt, Matcha. The evaluation results are shown in Table 16.

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. In contrast to the results of 2022 Matcha performs the matching task almost as quickly as LogMap, which stood out for its very fast calculation time. Apart from the runtime, the results for LogMap and LogMapLt do not change in comparison to the evaluation in 2022. LogMap presents a maximum precision value 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 to LogMap, LogMapLt calculates the alignment in around half the time and achieves much lower precision (0.4) but due to a greater amount of correctly found correspondences, the F1-measure is better - although still low with 0.143. Matcha finds 8 incorrect correspondences and thus is the worst-performing participant in terms of precision in the first test case. Investigating the reason for this low precision, Matcha appears to match classes with object

Table 15
Results for the Biodiversity & Ecology track.

System	Time (HH:MM:SS)	Number of mappings	Precision	Recall	F-measure
ENVO-SWEET task					
LogMap	00:00:36	681	0.780	0.655	0.713
LogMapKG	00:00:28	677	0.781	0.657	0.714
LogMapLt	00:05:40	595	0.829	0.594	0.692
OLaLa	06:46:18	1081	0.484	0.650	0.555
MACROALGAE-MACROZOOBENTHOS task					
LogMap	00:00:03	29	0.275	0.444	0.340
LogMapKG	00:00:04	29	0.275	0.444	0.340
LogMapLt	00:00:00	9	0.857	0.333	0.480
OLaLa	00:08:30	10	0.7	0.388	0.5
Matcha	00:00:07	45	0.2	0.5	0.285
FISH-ZOOPLANKTON task					
LogMap	00:00:03	32	0.093	0.2	0.127
LogMapKG	00:00:04	55	0.218	0.8	0.342
LogMapLt	00:00:00	8	1.0	0.533	0.695
OLaLa	00:07:59	13	1.0	0.866	0.928
Matcha	00:00:11	47	0.276	0.866	0.419
NCBITAXON-TAXREFLD Animalia task					
LogMap	00:00:43	72899	0.660	0.998	0.795
LogMapKG	00:11:32	72898	0.660	0.998	0.795
LogMapLt	00:00:43	72010	0.665	0.993	0.796
OLaLa	68:27:32	70821	0.679	0.998	0.808
Matcha	00:04:18	71008	0.674	0.993	0.803
NCBITAXON-TAXREFLD Bacteria task					
LogMap	00:00:01	304	0.575	1.0	0.730
LogMapKG	00:00:01	304	0.575	1.0	0.730
LogMapLt	00:00:00	290	0.6	0.994	0.748
OLaLa	00:19:32	294	0.593	0.994	0.743
Matcha	00:00:14	300	0.58	0.994	0.732
NCBITAXON-TAXREFLD Chromista task					
LogMap	00:00:04	2218	0.623	0.985	0.764
LogMapKG	00:00:01	2218	0.623	0.985	0.764
LogMapLt	00:00:01	2165	0.637	0.982	0.773
OLaLa	01:59:05	2173	0.634	0.981	0.771
Matcha	00:00:48	2213	0.624	0.984	0.764
NCBITAXON-TAXREFLD Fungi task					
LogMap	00:00:39	12949	0.783	0.998	0.878
LogMapKG	00:00:40	12949	0.783	0.998	0.878
LogMapLt	00:00:07	12929	0.783	0.997	0.877
OLaLa	11:54:37	12549	0.807	0.996	0.891
Matcha	00:01:43	12925	0.785	0.998	0.879
NCBITAXON-TAXREFLD Plantae task					
LogMap	00:01:44	26912	0.731	0.988	0.840
LogMapKG	00:01:36	26910	0.731	0.988	0.840
LogMapLt	00:00:17	26359	0.746	0.987	0.849
OLaLa	22:04:48	25667	0.769	0.991	0.866
Matcha	00:03:16	26597	0.741	0.989	0.847
NCBITAXON-TAXREFLD Protozoa task					
LogMap	00:00:01	496	0.719	1.0	0.837
LogMapKG	00:00:01	496	0.719	1.0	0.837
LogMapLt	00:00:00	477	0.746	0.997	0.853
OLaLa	00:32:17	476	0.750	1.0	0.857
Matcha	00:00:44	493	0.724	1.0	0.840

Table 16
Results for the three test cases of the MSE track.

System	Corresp.	Precision	Recall	F1-Measure	Time [s]
First Test Case					
LogMap	1	1.000	0.043	0.083	20
LogMapLt	5	0.400	0.087	0.143	53
Matcha	11	0.273	0.130	0.176	28
Second Test Case					
LogMap	67	0.881	0.195	0.320	6
LogMapLt	67	0.851	0.189	0.309	77
Matcha	87	0.756	0.219	0.339	15
Third Test Case					
LogMap	56	0.946	0.841	0.891	25
LogMapLt	56	0.911	0.810	0.857	53
Matcha	59	0.949	0.889	0.918	23

properties, e.g. “Temperature” = “hasTemperature” as in 2022. Since the recall is the best of the participating systems, Matcha turns out to be the best-performing system based on its F1-measure.

The second test case evaluates the matching systems to find correspondences between the large-sized MaterialInformation and the mid-sized BFO-based MatOnto. Surprisingly, LogMap performs the matching task significantly quicker than in the first test case and stands out for its very fast computation time of only 6s at a high precision of 0.881. Apart from the runtime, the results for LogMap and LogMapLt do not change in comparison to the evaluation in 2022. 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 significantly 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. Matcha presents the lowest precision in this test case but the highest recall. Based on its F1-measure, Matcha performs slightly better than the other systems in this 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 one minute. Apart from the runtime, the results for LogMap and LogMapLt do not change in comparison to the evaluation in 2022. All of the systems present high precision values. LogMap computes 3 false positives, LogMapLt computes 5 false positives and Matcha 3 false positives. At the same time Matcha computes with 56 the highest number of true positives, which results in the best precision of the tested systems. Since it has the highest number of correctly found correspondences, the recall and the F1-measure are the highest of the evaluated systems in the third test case at the fastest computation time.

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 LogMapLt is higher than for LogMap. In our opinion, LogMap is definitely recommended for MSE applications where high precision is demanded. In comparison to that, LogMapLt 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 7 systems that were able to finish the task within the 24-hour time limit with a non-empty alignment file: LogMap, OLaLa, Matcha, LogMapLt, LogMapKG, LsMatch, and AMD.

Table 17 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, OLaLa, Matcha, 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 OLaLa, Matcha, and AMD were able to generate alignments with the original dataset. Other systems either fail to complete the task within the allocated 24-hour time limit such as LogMapLt and LsMatch, or produce an empty alignment file such as LogMap (only on the YagoWikidata dataset). LogMapKG on the other hand tends to only align instances when it is applied to full-size datasets. Similar to the 2022 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[33]. On the Nell-DBpedia dataset, all systems were able to outperform the basic string matcher, in terms of f-measure, except for LogMapLt. On the YagoWikidata dataset, two systems were not able to outperform the baseline, which are LogMapLt and LsMatch. This year saw the return of different matchers and the introduction of a new one, OLaLa. While most

[33] https://oaei.ontologymatching.org/2023/results/commonKG/index.html

Table 17
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
OLaLa	120	1.0	0.92	0.96	00:07:07	original
LogMapLt	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
Matcha	114	0.99	0.87	0.93	00:01:53	original
String Baseline	78	1.00	0.60	0.75	00:00:37	original
Yago-Wikidata						
LogMap	233	1.00	0.76	0.86	00:00:26	small
OLaLa	209	1.0	0.68	0.81	00:03:56	original
LogMapLt	211	1.00	0.70	0.81	00:54:13	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	0.97	0.63	0.76	00:02:33	small
Matcha	274	0.99	0.90	0.94	00:07:16	original
String Baseline	212	1.00	0.70	0.82	00:00:02	original

matchers demonstrated similar performance to previous evaluations, Matcha notably improved its results on both datasets. Matcha also showcased the ability to function with the original datasets, a capability it lacked in the 2022 evaluation. OLaLa outperformed all other matchers in the Nell-DBpedia task, whereas Matcha excelled on the larger dataset, Yago-Wikidata. Furthermore, all matching processes were completed in less than an hour, as indicated in the runtime column. Lastly, the dataset size column specifies whether a system operated on the original dataset or solely on the smaller version.

4.12. Crosswalks Data Schema Matching

All the systems registered to OAEI 2023 were run besides the fact that only LogMap has been registered to participate in all tracks and no system has been specifically registered to the Crosswalks task.

This year, as introduced above, we have used the schemes for which an 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 schemes are not usually exposed in such a structured format. This opens the possibility of providing a dedicated task next year.

Table 18 shows the results for the systems that have generated correspondences. While generating a few number of correct correspondences, precision is higher with respect to recall for

Table 18
Results for the Crosswalks task.

	Correct	Output	Expected	Precision	Recall
LogMap					
datacity	0	3	34	0.00	0.00
iso	0	4	42	0.00	0.00
dcat3	1	12	42	0.08	0.02
dcterms	0	1	32	0.00	0.00
dcat-ap	0	2	34	0.00	0.00
	1	22	184	0.02	0.00
LogMapBio					
datacity	0	0	34	0.00	0.00
iso	0	0	42	0.00	0.00
dcat3	1	12	42	0.08	0.02
dcterms	1	1	32	1.00	0.03
dcat-ap	0	2	34	0.00	0.00
	2	15	184	0.22	0.01
LogMapKG					
datacity	0	0	34	0.00	0.00
iso	0	4	42	0.00	0.00
dcat	3 1	12	42	0.08	0.02
dcterms	0	1	32	0.00	0.00
dcat-ap	0	2	34	0.00	0.00
	1	19	184	0.02	0.00
LogMapLt					
datacity	0	4	34	0.00	0.00
iso	1	14	42	0.07	0.02
dcat3	4	41	42	0.10	0.10
dcterms	1	3	32	0.33	0.03
dcat-ap	0	4	34	0.00	0.00
	6	66	184	0.10	0.03
Matcha					
datacity	0	5	34	0.00	0.00
iso	1	10	42	0.10	0.02
dcat3	4	44	42	0.09	0.10
dcterms	1	1	32	1.00	0.03
dcat-ap	0	0	34	0.00	0.00
	6	60	184	0.24	0.03

all systems. Most of the generated correspondences still involve properties where labels are equal, for instance: https://schema.org/distribution and http://www.w3.org/ns/dcat#distribution. In terms of F-measure, Matcha and LogMapLt have the best and similar performance. With respect to the pairs, a higher number of correspondences has been generated for the pairs involving DCAT-v3. LogMapLt and Matcha are the systems that are able to deal with a higher number of matching pairs.

In 2022, this track ran for the first time. Last year, similar to this year, only Matcha, LogMap and LogMapLt were able to generate non-empty alignments, with LogMapLt being able to generate a higher number of correspondences. In terms of precision, Matcha and LogMapLt had a higher precision in detriment of recall.

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

4.13. Knowledge Graph

This year we evaluated all participants with the MELT framework to include all possible submission formats i.e. SEALS, and Web format. First, all systems are evaluated on a very small matching task[34] (even those not registered for the track). This revealed that not all systems were able to handle the task, and in the end, 6 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 contain 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 [35].

Table 19 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 distinguish empty and erroneous (or not generated) alignments. The values in parentheses show the results when considering only nonempty alignments.

This year's best overall system is the baseline using the alternative labels (0.84 F-measure). The highest recall is again achieved by Matcha (0.84). It returns more correspondences than all others (263,822.2 on average) but is only able to match instances in this track. Detailed results for each test case can be found on the OAEI results page of the track[36].

Property matches are still not created by all systems. LogMap, Matcha, and SORBETMatcher 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`. OLaLa reaches the best score with 0.83 F-Measure.

Regarding runtime, Matcha (14:30:03) and LogMapLt (64:48:07) were the slowest systems. In comparison to last year, the runtimes increased quite a lot and the systems should focus more on scalable solutions. Besides the baselines (which need around 12 minutes for all test cases) LogMap (00:56:43) and SORBETMatcher (00:21:53) were the fastest systems.

For further analysis of the results, we also provide an online dashboard[37] generated with MELT[65]. It allows us 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.

4.14. Pharmacogenomics

For this first year of the Pharmacogenomics track, 2 systems registered, namely LogMap and Matcha. The evaluation was performed using the MELT framework. Unfortunately, none

[34] http://oaei.ontologymatching.org/2019/results/knowledgegraph/small_test.zip
[35] http://oaei.ontologymatching.org/2023/results/knowledgegraph/knowledgegraph-alignments.zip
[36] http://oaei.ontologymatching.org/2023/results/knowledgegraph/index.html
[37] http://oaei.ontologymatching.org/2023/results/knowledgegraph/knowledge_graph_dashboard.html

Table 19
Knowledge Graph track results, divided into class, property, and instance performance. For matchers that were not capable of completing all tasks, the numbers in parentheses denote the performance when only averaging across tasks that were completed.

System	Time	tracks	Size	Prec.	F-m.	Rec.
class performance						
BaselineAltLabel	00:11:37	5	16.4	1.00 (1.00)	0.71 (0.71)	0.59 (0.59)
BaselineLabel	00:11:27	5	16.4	1.00 (1.00)	0.71 (0.71)	0.59 (0.59)
LogMap	00:56:43	5	19.4	0.93 (0.93)	0.80 (0.80)	0.71 (0.71)
LogMapLt	64:48:07	4	23.0	0.80 (1.00)	0.55 (0.69)	0.43 (0.54)
LSMatch	04:47:07	5	23.6	0.97 (0.97)	0.74 (0.74)	0.64 (0.64)
Matcha	14:30:03	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
OLaLa	02:55:06	5	18.6	0.98 (0.98)	0.68 (0.68)	0.53 (0.53)
SORBETMatcher	00:21:53	5	22.4	0.93 (0.93)	0.80 (0.80)	0.73 (0.73)
property performance						
BaselineAltLabel	00:11:37	5	47.8	0.99 (0.99)	0.76 (0.76)	0.66 (0.66)
BaselineLabel	00:11:27	5	47.8	0.99 (0.99)	0.76 (0.76)	0.66 (0.66)
LogMap	00:56:43	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
LogMapLt	64:48:07	4	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
LSMatch	04:47:07	5	85.6	0.73 (0.73)	0.71 (0.71)	0.69 (0.69)
Matcha	14:30:03	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
OLaLa	02:55:06	5	73.6	0.86 (0.86)	0.83 (0.83)	0.81 (0.81)
SORBETMatcher	00:21:53	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
instance performance						
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)	0.80 (0.80)	0.71 (0.71)
LogMap	00:56:43	5	4012.4	0.90 (0.90)	0.78 (0.78)	0.69 (0.69)
LogMapLt	64:48:07	4	6653.8	0.73 (0.91)	0.67 (0.84)	0.62 (0.78)
LSMatch	04:47:07	5	5872.2	0.66 (0.66)	0.59 (0.59)	0.60 (0.60)
Matcha	14:30:03	5	263822.2	0.55 (0.55)	0.63 (0.63)	0.86 (0.86)
OLaLa	02:55:06	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
SORBETMatcher	00:21:53	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
overall performance						
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)	0.80 (0.80)	0.71 (0.71)
LogMap	00:56:43	5	4031.8	0.90 (0.90)	0.77 (0.77)	0.68 (0.68)
LogMapLt	64:48:07	4	6676.8	0.73 (0.92)	0.66 (0.83)	0.61 (0.76)
LSMatch	04:47:07	5	5981.4	0.66 (0.66)	0.60 (0.60)	0.61 (0.61)
Matcha	14:30:03	5	263822.2	0.55 (0.55)	0.62 (0.62)	0.84 (0.84)
OLaLa	02:55:06	5	92.2	0.88 (0.88)	0.03 (0.03)	0.02 (0.02)
SORBETMatcher	00:21:53	5	22.4	0.93 (0.93)	0.01 (0.01)	0.00 (0.00)

of these systems successfully produced alignments between reified n-ary tuples representing pharmacogenomic knowledge units. Some configurations of the two systems output some alignments between other entities (*e.g.*, components of pharmacogenomic tuples) but not between the n-ary tuples themselves.

5. Conclusions and Lessons Learned

As in previous campaigns, in 2023, we witnessed a healthy mix of new and returning systems, with an imbalanced participation in the tracks.

The **schema matching** tracks gather the highest number of participants; however still little substantial progress in terms of the quality of the results or run time of top matching systems. As already reported last year, we observe a performance plateau being reached by existing strategies and algorithms. 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 are more systems with the ability to match properties (7 in 2023 vs. 5 in 2022). Several ML-based systems used reference alignments for training to a certain extent (explained in their system papers). It has already happened in the past. It calls for a discussion and perhaps more ML-based tracks.

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 Matportal[38]. 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[39], the MSEO[40], the PMDco[41], the prov-o[42] and IOF-mat[43]. In the OntoCommons-project[44] alignments of frequently used ontologies of the MSE application area are produced and will be used to further improve the MSE benchmark based on the project results. Apart from also considering frequently used domain and application ontologies, also multi-ontology matching, knowledge graph matching, e.g. using the AluTrace-data[45], and the usage of background knowledge should be considered in future OAEI campaigns.

With respect to the cross-lingual version of the Conference, the **Multifarm** track still attracts too few number of participants. Despite this fact, this year new participants came up with alternative strategies (i.e., deep learning) with respect to the last campaigns.

In the **Food** track, none of the evaluated matchers finds all reference correspondences correctly. LogMapLt stands out for its very fast computing speed. Matcha obtains the best results for the FNC application. 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 incorporated significant updates and attracted several new machine learning-based participants. However, the number of symbolic participants decreased. The best-performing systems are not consistent across tasks and settings, demonstrating the diversity of our datasets. It is also worth noting that SORBETMatcher is the only system can participate in both equivalence

[38] https://matportal.org/
[39] https://matportal.org/ontologies/BWMD-MID
[40] https://matportal.org/ontologies/MSEO
[41] https://github.com/materialdigital/core-ontology/
[42] https://www.ebi.ac.uk/ols/ontologies/prov
[43] https://industrialontologies.org/working-groups/the-material-science-and-engineering-mse-working-group-wg/
[44] https://ontocommons.eu
[45] https://github.com/Mat-O-Lab/AluTraceProject

and subsumption matching.

In the **Biodiversity and Ecology** track, none of the systems was able to detect manual mappings created by domain experts and requiring biodiversity domain-specific knowledge. In this year's edition, we confirmed the inability of most systems to handle SKOS natively, as well as very large ontologies. Additionally, some systems did not perform well on the thesauri tasks because those contained concepts with similar URIs that were, in fact, completely different.

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 9 and 11 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 attracts too few number of participants. This year, no system was able to complete the task. As several sub-tracks have been discontinued, the track is limited to the conference domain. This track welcomes new organizers.

The **Crosswalks Data Schema Matching** track involves different schema formats and ways of representing schema properties. This opens the possibility of creating a dedicated task relying on other formats than OWL/RDF.

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, 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.

In 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, with a new system participating and former systems adapting their approaches to scale up to the task size. However, with some systems only being able to produce alignments when applied to smaller versions of the KG datasets, we still look forward to having more participants in the next OAEI campaign.

For the first year of the **Pharmacogenomics** track, participation was limited with only 2 systems registered. Unfortunately, none of the participating systems were able to output alignments between the targeted reified n-ary tuples. We will investigate whether this originates from the absence of labels for tuples and the only presence of structural information or if other aspects are detrimental (*e.g.*, arity, background domain knowledge that must be considered to produce most alignments). These results highlight the interest in considering domain-specific problems to design new methods like [56, 66] or enrich existing ones. Recall that the track features different types of alignments between individuals, which is a specificity of the considered alignment task. This raises the question of whether such a granular matching setting could be generalized to other instance matching tasks. Since the alignment task in this track is structure-based, it is particularly

well-adapted to approaches relying on Graph Neural Networks that learn embeddings of nodes to align on the basis of their neighborhoods [66]. All these reasons motivate to propose again the track in the next editions of OAEI and adapt it to evaluate Machine Learning-based matchers. We hope that the growing awareness about this track and its specificity will attract additional systems.

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 Workshop point out different directions for future improvements in OAEI. This year, with a higher number of systems relying on Large Language Models, there was a discussion on the specific requirements and alternative ways for gathering the alignments generated by such resource-consuming systems. It has also been highlighted the need to push the adoption of SSSOM [24] (this year MELT has incorporated the format but still few systems have adopted it), as a way for delivering richer alignments in terms of metadata and justifications [67]. As already mentioned before, there were also some interrogations on the stability reached in some (open)-schema matching tasks (in particular Anatomy and Conference tracks) as the performance has been quite stable for several years. This requires a further analysis of the difficult parts of the matching task. Last but not least, new tracks addressing more application/use-oriented tasks should be addressed and they are more than welcome.

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.

Acknowledgments

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 in improving the dataset's quality.

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), and 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 program 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) 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).

Naouel Karam and Alsayed Algergawy have been supported by the German Research Foundation in the context of NFDI4BioDiversity project (number 442032008) and the CRC 1076 AquaDiva. We would like to thank Jessica Titocci, Martina Pulieri and Ilaria Rosati for providing the datasets for the biodiv SKOS thesauri tasks.

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, P. Buche, L. J. Castro, J. Chen, H. Dong, O. Fallatah, D. Faria, I. Fundulaki, S. Hertling, Y. He, I. Horrocks, 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, T. Saveta, P. Shvaiko, C. Trojahn, C. Verhey, M. Wu, B. Yaman, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2022, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn (Eds.), Proceedings of the 17th International Workshop on Ontology Matching (OM 2022) co-located with the 21th International Semantic Web Conference (ISWC 2022), Hangzhou, China, held as a virtual conference, October 23, 2022, volume 3324 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 84–128. URL: https://ceur-ws.org/Vol-3324/oaei22_paper0.pdf.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] N. Matentzoglu, J. P. Balhoff, S. M. Bello, C. Bizon, M. Brush, T. J. Callahan, C. G. Chute, W. D. Duncan, C. T. Evelo, D. Gabriel, J. Graybeal, A. Gray, B. M. Gyori, M. Haendel, H. Harmse, N. L. Harris, I. Harrow, H. B. Hegde, A. L. Hoyt, C. T. Hoyt, D. Jiao, E. Jiménez-Ruiz, S. Jupp, H. Kim, S. Koehler, T. Liener, Q. Long, J. Malone, J. A. McLaughlin, J. A. McMurry, S. Moxon, M. C. Munoz-Torres, D. Osumi-Sutherland, J. A. Overton, B. Peters, T. Putman, N. Queralt-Rosinach, K. Shefchek, H. Solbrig, A. Thessen, T. Tudorache, N. Vasilevsky, A. H. Wagner, C. J. Mungall, A Simple Standard for Sharing Ontological Mappings (SSSOM), Database 2022 (2022) baac035. URL: https://doi.org/10.1093/database/baac035. doi:10.1093/database/baac035.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] N. A. Vasilevsky, N. A. Matentzoglu, S. Toro, J. E. Flack IV, H. Hegde, D. R. Unni, G. F. Alyea, J. S. Amberger, L. Babb, J. P. Balhoff, et al., Mondo: Unifying diseases for the world, by the world, medRxiv (2022) 2022–04.

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

[36] E. Jiménez-Ruiz, B. C. Grau, LogMap: Logic-based and scalable ontology matching, in: Proceedings of the 10th International Semantic Web Conference, Bonn (DE), 2011, pp. 273–288.

[37] Y. He, J. Chen, H. Dong, I. Horrocks, Exploring large language models for ontology alignment, arXiv preprint arXiv:2309.07172 (2023).

[38] Y. He, J. Chen, H. Dong, I. Horrocks, C. Allocca, T. Kim, B. Sapkota, Deeponto: A python package for ontology engineering with deep learning, arXiv preprint arXiv:2307.03067 (2023).

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] A. Algergawy, N. Karam, A. Laadhar, F. Michel, Too big to match: a strategy around matching tasks for large taxonomies, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn (Eds.), Proceedings of the 17th International Workshop on Ontology Matching (OM 2022) co-located with the 21th International Semantic Web Conference (ISWC 2022), Hangzhou, China, held as a virtual conference, October 23, 2022, volume 3324 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 67–72. URL: https://ceur-ws.org/Vol-3324/om2022_STpaper1.pdf.

[44] T. Ashino, Materials Ontology: An Infrastructure for Exchanging Materials Information

and Knowledge, Data Science Journal 9 (2010) 54–61.
[45] 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.
[46] 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.
[47] 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.
[48] 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.
[49] 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.
[50] 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.
[51] 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.
[52] P. Krauss, schemaorg-wikidata-map, https://github.com/okfn-brasil/schemaOrg-Wikidata-Map, 2017.
[53] S. Hertling, H. Paulheim, Dbkwik: extracting and integrating knowledge from thousands of wikis, Knowledge and Information Systems (2019).
[54] S. Hertling, H. Paulheim, Dbkwik: A consolidated knowledge graph from thousands of wikis, in: Proceedings of the International Conference on Big Knowledge, 2018.
[55] P. Monnin, A. Coulet, Matching pharmacogenomic knowledge: particularities, results, and perspectives, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn (Eds.), Proceedings of the 17th International Workshop on Ontology Matching (OM 2022) co-located with the 21th International Semantic Web Conference (ISWC 2022), Hangzhou, China, held as a virtual conference, October 23, 2022, volume 3324 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 79–83. URL: https://ceur-ws.org/Vol-3324/om2022_STpaper3.pdf.
[56] P. Monnin, M. Couceiro, A. Napoli, A. Coulet, Knowledge-based matching of n-ary tuples, in: M. Alam, T. Braun, B. Yun (Eds.), Ontologies and Concepts in Mind and Machine - 25th International Conference on Conceptual Structures, ICCS 2020, Bolzano, Italy, September 18-20, 2020, Proceedings, volume 12277 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 48–56. URL: https://doi.org/10.1007/978-3-030-57855-8_4. doi:10.1007/978-3-030-57855-8_4.
[57] P. Monnin, J. Legrand, G. Husson, P. Ringot, A. Tchechmedjiev, C. Jonquet, A. Napoli, A. Coulet, PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison, BMC Bioinformatics 20-S (2019) 139:1–139:16. URL: https://doi.org/10.1186/s12859-019-2693-9. doi:10.1186/

S12859-019-2693-9.
[58] 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.
[59] J. Chen, Y. He, Y. Geng, E. Jiménez-Ruiz, H. Dong, I. Horrocks, Contextual semantic embeddings for ontology subsumption prediction, World Wide Web (2023) 1–23.
[60] 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.
[61] Z. Wang, I. F. Cruz, Agreementmakerdeep results for oaei 2021., in: OM@ ISWC, 2021, pp. 124–130.
[62] D. Faria, M. C. Silva, P. Cotovio, P. Eugénio, C. Pesquita, Matcha and matcha-dl results for oaei 2022, 2022.
[63] F. Gosselin, A. Zouaq, Sorbet: A siamese network for ontology embeddings using a distance-based regression loss and bert, in: International Semantic Web Conference, Springer, 2023, pp. 561–578.
[64] 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.
[65] 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.
[66] P. Monnin, C. Raïssi, A. Napoli, A. Coulet, Discovering alignment relations with graph convolutional networks: A biomedical case study, Semantic Web 13 (2022) 379–398. URL: https://doi.org/10.3233/SW-210452. doi:10.3233/SW-210452.
[67] N. Matentzoglu, I. Braun, A. R. Caron, D. Goutte-Gattat, B. M. Gyori, N. L. Harris, E. Hartley, H. B. Hegde, S. Hertling, C. Tapley, H. Kim, H. Li, J. McLaughlin, C. Trojahn, N. Vasilevsky, C. Mungall, A Simple Standard for Ontological Mappings 2023: Updates on data model, collaborations and tooling, in: OM 2023: The 18th International Workshop on Ontology Matching collocated with the 22nd International Semantic Web Conference ISWC-2023 November 7th, 2023, Athens, Greece, 2023.

Linköping, Jena, Lisboa, Heraklion, Mannheim, Montpellier, Oslo, London, Berlin, Trento, Toulouse, Prague, Manhattan, Dublin, Oxford
December 2023