Results of the Ontology Alignment Evaluation Initiative 2024

Mina Abd Nikooie Pour[1,2], Alsayed Algergawy[3,4], Eva Blomqvist[1], Patrice Buche[5], Jiaoyan Chen[6], Pedro Giesteira Cotovio[7,8], Adrien Coulet[9,10], Julien Cufi[5], Hang Dong[11], Daniel Faria[12], Lucas Ferraz[8], Sven Hertling[13], Yuan He[14], Ian Horrocks[14], Liliana Ibanescu[15], Sarika Jain[16], Ernesto Jiménez-Ruiz[7], Naouel Karam[17], Felix Kraus[18], Patrick Lambrix[1,2], Huanyu Li[1], Ying Li[1,2], Pierre Monnin[19], Heiko Paulheim[13], Catia Pesquita[8], Abhisek Sharma[16], Pavel Shvaiko[20], Marta Silva[8], Guilherme Sousa[21], Cassia Trojahn[22], Jana Vataščinová[23], Beyza Yaman[24], Ondřej Zamazal[23] and Lu Zhou[25]

[1] *Department of Computer and Information Science, Linköping University, Linköping, Sweden*
[2] *Swedish e-Science Research Centre, Linköping, Sweden*
[3] *Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Germany*
[4] *Chair of Data and Knowledge Engineering, University of Passau, Germany*
[5] *UMR IATE, INRAE, University of Montpellier, France*
[6] *Department of Computer Science, The University of Manchester, UK*
[7] *City St George's, University of London, UK & University of Oslo, Norway*
[8] *LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[9] *Inria Paris, France*
[10] *Centre de Recherche des Cordeliers, Inserm, Université Paris Cité, Sorbonne Université, France*
[11] *Department of Computer Science, University of Exeter, UK*
[12] *INESC-ID / IST, University of Lisbon, Portugal*
[13] *Data and Web Science Group, University of Mannheim, Germany*
[14] *Department of Computer Science, University of Oxford, UK*
[15] *Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA Paris-Saclay, France*
[16] *National Institute of Technology Kurukshetra, Haryana, India*
[17] *Institute for Applied Informatics, University of Leipzig, Germany*
[18] *Karlsruhe Institute of Technology, Karlsruhe, Germany*
[19] *Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France*
[20] *Trentino Digitale SpA, Trento, Italy*
[21] *Institut de Recherche en Informatique de Toulouse, France*
[22] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France*
[23] *Prague University of Economics and Business, Czech Republic*
[24] *ADAPT Centre, Trinity College Dublin*
[25] *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 2024 campaign offered 13 tracks and was attended by 13 participants. This paper is an overall presentation of that campaign.

OM 2024: The 19th International Workshop on Ontology Matching collocated with the 23rd International Semantic Web Conference (ISWC 2024), November 11th, 2024, Baltimore, USA

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 has been run for 20 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.

The first two 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, 22], which this year took place in Baltimore, USA[2].

Since 2011, we have been using an environment for automatically processing evaluations that was developed within the SEALS (Semantic Evaluation At Large Scale) project[3]. SEALS provided a software infrastructure to automatically execute 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 2022, the MELT framework [23] has been adopted to facilitate the SEALS and HOBBIT wrapping and evaluation. Since 2023, most tracks have adopted MELT as their evaluation platform.

This paper synthesizes the 2024 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.

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 [24].

The **MELT framework**[5] [23] was introduced in 2019 and is under active development. It allows the development, evaluation, and packaging of matching systems for evaluation interfaces like SEALS

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

[2] http://om.ontologymatching.org/2024

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

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

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

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 [25] 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, the evaluation was carried out *a posteriori*, using the alignments produced by the matching systems.

2.2. Submission formats

As already mentioned above, three submission formats were allowed: (1) SEALS package, (2) HOBBIT, and (3) MELT. 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 a docker container that implements the HTTP interface.

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.

2.3. OAEI campaign phases

As in previous years, the OAEI 2024 campaign was divided into three phases: preparatory, execution, and evaluation.

In the **preparation phase**, the test cases were provided to participants during an initial evaluation period between June 30^{th} and July 31^{st}, 2024. The goal of this phase is to ensure that the test cases make sense to the 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 subsequent **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}, 2024, at which date participants had to submit the (near) final versions of their systems (SEALS-wrapped and/or HOBBIT-wrapped).

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

3. Tracks and Test Cases

This year's OAEI campaign consisted of 13 tracks, all of them including OWL ontologies while only one also including SKOS thesauri. They can be grouped into:

- Schema matching tracks, which have as objective matching ontology classes and/or properties.

[6] https://dwslab.github.io/melt/matcher-evaluation/client

[7] https://dwslab.github.io/melt/matcher-packaging/web

- 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 2024.

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			√
circular economy	OWL	=	[0 1]	open	EN			√
dh	SKOS	=	[0 1]	open	AR, DE, EN, ES, FR, HR, HU, IT, NL, SL			√
arch-multiling	SKOS	=	[0 1]	open	DE, EN, FR, IT			√
Instance and schema matching								
knowledge graph	OWL	=	[0 1]	open	EN			√
Instance matching or link discovery								
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 [26].

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.

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

3.2. Conference

The conference track consists of a suite of 21 matching tasks corresponding to the pairwise combination of 7 moderately expressive ontologies describing the domain of organizing conferences. The dataset and its usage is described in [27].

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 [28] 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 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.

3.4. 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*.

As last 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 participants of the track output their (complex) correspondences in the EDOAL format. For the Conference dataset, the complex correspondences are manually compared to the ones of the consensus alignment. For the Populated Conference dataset, the alignments are evaluated using coverage and precision metrics using an evaluator that relies on the comparison of sets of instances [29]. All our evaluations were conducted on a server machine with AMD EPYC 7402 2.8 GHz x48 processors, 512GB RAM. Processes needing a GPU were run in a compute node with 4 Nvidia Geforce GTX 1080TI.

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 [30], 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 [31, 32, 33]. The evaluation thus focuses on how interaction with the user improves the matching results. Currently, this track does not evaluate the user experience or the user interfaces of the systems [34, 32].

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

[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

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 [35] incorporates both *equivalence* and *subsumption* ontology matching (OM) tasks for biomedical ontologies, with ground truth (equivalence) mappings extracted from Mondo [36] and UMLS [37] (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 [35], we make available 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 addition, the Bio-LLM sub-track supports a more efficient evaluation of large language model-based OM [38], which consists of challenging subsets of NCIT-DOID and SNOMED-FMA (Body) datasets, along with tailored evaluation metrics. Since the 2023 edition, Bio-ML has added a *logical module enrichment* [39] to add entities to the pruned ontologies to provide more context for alignment, annotated as *"not used in alignment"* and ignored in evaluation. For evaluation, in [35] 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, 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($\equiv$)** and **#Refs ($\sqsubseteq$)**, respectively.

Mapping Source	Ontology Pair	Category	#Refs ($\equiv$)	#Refs ($\sqsubseteq$)
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 that the results came from MELT, source code implementation, and direct result submission, respectively. All our evaluations were conducted with the DeepOnto[18] [40] library.

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 [41, 42, 43]. 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 [44]. 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 [45]. 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. Digital Humanities

The use of controlled vocabularies is widespread within the digital humanities (DH) [46]. The development and usage of these vocabularies by different parties in related domains naturally leads to overlaps in content [47]. While ontology matching helps with alignment and integration tasks, the application of these systems to the digital humanities poses special challenges. Highly specific domain terminology often leads to smaller vocabularies, which oftentimes include multiple (ancient) languages. Furthermore, matching systems need to be compatible with SKOS vocabularies, since their use is fairly common within the community.

The DH track participated for the first time this year. It includes eight test cases from archaeology, cultural history and DH / computer science. Each test case consists of two SKOS (using RDF/XML as

[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.
[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	-

syntax) vocabularies to be matched and one manually created gold standard reference. For details on the nine source vocabularies and on the test cases, see Table 5 and Table 6.

The evaluation was executed on a virtual machine with 8 cores (2.4GHz each) and 16 GB RAM. To quantify the performance, precision, recall and F1-score were used, while only evaluating equivalence relationships. If matching systems resulted in either errors or zero identified matches, the task was considered as failed. Adhering to the OAEI rules, no settings were changed before running the matching systems.

Table 5
Controlled vocabularies used for the digital humanities (dh) track.

Resource	Field[24]	Version / Date	#concepts[25]	language (ISO 639)
DEFC Thesaurus[26]	Archaeology	-	∼800	de, en, la
PACTOLS thesaurus for archaeology[27]	Archaeology	- / 2021-05-18	∼60,000	ar, de, en, es, fr, it, nl
Iron-Age-Danube thesaurus[28]	Archaeology	1 / 2018-11-07	∼6900	de, en, hr, hu, sl
iDAI.world Thesaurus[29]	Arch. / cult. hist.	1.2 / 2022-02-10	∼290	de, en, es, fr, it
PARTHENOS Vocabularies[30]	Arch. / cult. hist.	- / 2019-05-07	∼4200	en
OeAI Thesaurus - Cultural Time Periods[31]	Cultural history	1.0.0 / 2022-11-23	∼400	de, en
DHA Taxonomy[32]	DH/CS	- / 2018-04-03	∼120	en
UNESCO[33]	DH/CS	- / 2024-06-03	∼4500	ar, en, fr, es, ru
TaDiRAH[34]	DH/CS	2.0.1 / 2021-07-22	∼170	de, en, es, fr, it, pt, sr

[24] This is the field to which the CV was grouped within our dataset.

[25] This is the number of concepts in the primary language of the CV before any preprocessing steps.

[26] https://vocabs.dariah.eu/defc_thesaurus/en/

[27] https://isl.ics.forth.gr/bbt-federated-thesaurus/PACTOLS/en/

[28] https://vocabs.dariah.eu/iad_thesaurus/en/

[29] https://isl.ics.forth.gr/bbt-federated-thesaurus/DAI/en/

[30] https://vocabs.dariah.eu/parthenos_vocabularies/en/

[31] https://vocabs.acdh.oeaw.ac.at/oeai-cp/en/

[32] https://vocabs.dariah.eu/dha_taxonomy/en/

[33] https://vocabularies.unesco.org/browser/thesaurus/en/

[34] https://vocabs.dariah.eu/tadirah/en/

[35] The number of terms varies depending on the branch used for the respective domain.

Table 6
Properties of the digital humanities (dh) track.

Domain	Source (#terms[35])	Target (#terms)	#True Positives
Archaeology	DEFC (800)	PACTOLS (70)	11
	iDAI (2600)	PACTOLS (70)	18
	Iron-Age-Danube (290)	PACTOLS (70)	6
	PACTOLS (70)	PARTHENOS (800)	13
Cultural History	iDAI (270)	PARTHENOS (200)	53
	OeAI (400)	PARTHENOS (200)	48
DH / CS	DHA (115)	UNESCO (490)	12
	TaDiRAH (170)	UNESCO (490)	16

3.10. Archaeology multilingual

The archaeology multilingual track is based on an archaeology test case of the digital humanities track, see section 3.9, with focus on evaluating matcher performance when dealing with different languages.

Like the DH track, this track participated for the first time. Each test case uses iDAI.world and PACTOLS (see Table 5 for more information) as source resp. target. Both vocabularies contain terms in English, French, German, and Italian. To create the ten test cases, all but one language were removed from both vocabularies, leading to 10 different test cases, consisting of two monolingual vocabularies and a manually created gold standard reference.

The evaluation modalities are identical to ones in the digital humanities track, see section 3.9.

3.11. Circular Economy

In recent years, the Circular Economy (CE) domain has shown interest in representing domain knowledge using ontologies. Since there are some existing CE-specific ontologies with more emerging, providing alignments among ontologies can enhance the interoperability and reusability of such ontologies. Therefore the circular economy track is proposed in 2024 consisting 1 task to match 2 ontologies in the circular economy domain. These two ontologies are the Circular Economy Ontology Network [48] and the Sustainable Bioeconomy and Bioproducts Ontology (BiOnto) [49]. CEON (including 214 classes) from the Onto-DESIDE project,[36] aims to represent core concepts for the CE domain [48]. BiOnto (including 780 classes) from the BIOVOICES project,[37] focuses on establishing a shared and common terminology in the bioeconomy domain so that different stakeholders participating circular value networks can provide information according to the ontology [50].

The evaluation is conducted over standard parameters which are precision, recall, f-measure and alignment size. The reference alignment for the matching task was initially done in [51] and further validated by ontology engineers and CE domain experts from Onto-DESIDE project. The results is presented in Section 4.12.

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[38] in the course of the DBkWik project [52, 53]. 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

[36] https://ontodeside.eu
[37] https://www.biovoices.eu
[38] 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

concept, a few restrictions were made: 1) only links in sections with a header containing "link" are used, 2) all links are removed where the source page links to more than one concept in another wiki (ensures the alignments are functional), 3) multiple links which point to the same concept are also removed (ensures injectivity), 4) links to disambiguation pages were manually checked and corrected. Since we do not have a correspondence for each instance, class, and property in the graphs, this gold standard is only a *partial gold standard*.

The evaluation was executed on a virtual machine (VM) with 32GB of RAM and 16 vCPUs (2.4 GHz), with Debian 9 operating system and Openjdk version 1.8.0_265. For evaluating all possible submission formats, MELT framework is used. The corresponding code for evaluation can be found on Github[39].

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[40].

3.13. Pharmacogenomics

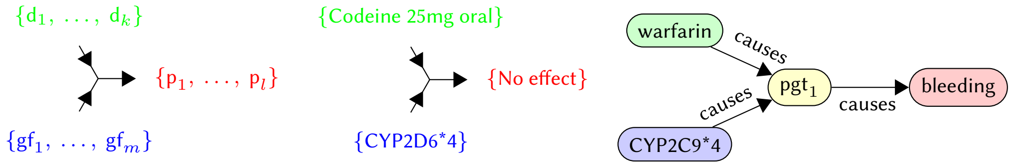
In 2024, the Pharmacogenomics track was run for the second time. This track focuses on matching knowledge units from the pharmacogenomics domain. These units are n-ary tuples – so-called "pharmacogenomic relationships" – and involve drugs, genetic factors, and phenotypes (see Figure 1). A pharmacogenomic 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 predicates (Figure 1(c)). Hence, the task of matching pharmacogenomic tuples is [54]:

- 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 the only available information about these tuples is their neighbors (*e.g.*, no labels, or other properties).

To illustrate, two tuples associating the same sets of drugs, genetic factors, and phenotypes have to the same neighbors, thus represent the same two "pharmacogenomics relationships", and thus should be detected as identical.

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



(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.

Beside the arity of tuples, matchers need to face issues such as 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 [54, 55] 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 the 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[41] [56]. 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.

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 13 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 [57] and BERTSubs [58]) that are not counted in the number of participants. 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,

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

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.

Table 8
Participants and the status of their submissions.

System	ALIN	BERTMap	BERTMapLt	BioSTransformers	CANARD	DeepLSMatch	HybridOM	LogMap	LogMap-Bio	LogMapLt	LogMapKG	Matcha	MDMapper	OntoMatch	TOMATO	Total=13
anatomy	●	○	○	○	○	○	○	●	●	●	○	●	●	○	●	7
conference	●	○	○	○	○	○	○	●	○	●	○	●	●	●	●	7
multifarm	○	○	○	○		○	○	●	○	●	○	●	●	○	○	4
complex	○	○	○	○	◐	○	○	○	○	○	○	○	○	○	○	1
food	○	○	○	○	○	○	○	●	○	●	○	●	○	○	○	3
interactive	●	○	○	○	○	○	○	●	○	○	○	○	○	○	○	2
bio-ML	○	●	●	●	○	●	●	●	●	●	○	●	○	○	○	9
biodiv	○	○	○	○	○	○	○	●	○	●	●	●	○	○	○	4
circular economy	○	○	○	○	○	○	○	●	○	●	○	●	○	○	○	3
dh	○	○	○	○	○	○	○	●	●	○	●	●	○	○	○	4
arch-multiling	○	○	○	○	○	○	○	●	●	○	●	●	○	○	○	4
knowledge graph	○	○	○	○	○	○	○	●	○	●	●	●	●	○	○	5
pharmacogenomics	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	0
total	3	1	1	1	1	1	1	11	4	8	4	10	4	1	2	

4.2. Anatomy

The results for the Anatomy track are shown in Table 9. Of the 7 systems participating in the Anatomy track, 6 achieved an F-measure higher than the StringEquiv baseline. Two systems were first-time participants (TOMATO, MDMapper) in anatomy track. Long-term participating systems showed few changes in comparison with previous years with respect to alignment quality (precision, recall, F-measure, and recall+) and size. The exception were Matcha which increased in size (from 1484 to 1485), recall+ (from 0.818 to 0.82), LogMapBio decreased in size (from 1578 to 1549), recall (from 0.916 to 0.908), recall+ (from 0.778 to 0.757), increased in precision (from 0.88 to 0.888) and ALIN decreased in size (from 1159 to 1156), F-measure (0.852 to 0.851), recall (from 0.752 to 0.75), recall+ (from 0.501 to 0.489). In terms of run time, 3 out of 7 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+ (0.82), but three other systems obtained an F-measure above 0.88 (LogMapBio, MDMapper, 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).

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.

With regard to two baselines we can group tools according to system's position: there are six matchers above (or equal to) edna baseline (ALIN, LogMap, LogMapLT, Matcha, MDMapper, and OntoMatch),

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	42	1485	0.951	0.941	0.931	0.82	-
LogMapBio	1346	1549	0.888	0.898	0.908	0.757	√
MDMapper	121	1441	0.926	0.903	0.881	0.703	-
LogMap	12	1402	0.917	0.881	0.848	0.602	√
ALIN	370	1156	0.984	0.851	0.75	0.489	-
LogMapLt	2	1147	0.962	0.828	0.728	0.288	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
TOMATO	2154	572	0.955	0.523	0.36	0.024	-

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.
LogMap	0.76	0.71	0.64	0.59	0.56	0	21	0
Matcha	0.66	0.65	0.64	0.64	0.63	7	86	72
MDMapper	0.66	0.63	0.59	0.55	0.53	2	29	13
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	3	97	18
OntoMatch	0.82	0.69	0.56	0.48	0.43	0	2	0
StringEquiv	0.76	0.65	0.53	0.45	0.41			
TOMATO	0.57	0.53	0.48	0.44	0.42	12	353	162

and one matcher below StringEquiv baseline (TOMATO). Three matchers (ALIN, MDMapper, and OntoMatch) do not match properties at all.

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 7 alignment systems, 5 (ALIN, LogMapLt, MDMapper, OntoMatch, TOMATO) use 1.0 as the confidence value for all matches they identify. The remaining 2 systems (LogMap, Matcha) have a wide variation of confidence values.

The evaluation results show key differences in how matchers handle uncertain reference alignments, particularly in discrete and continuous metrics.

ALIN and OntoMatch both maintain high precision (0.88) across metrics and significantly improve in recall and F-measure when moving from sharp to uncertain settings, demonstrating strong adaptability to uncertain alignments.

LogMap and LogMapLt perform consistently, with LogMap maintaining stable precision (0.81) and LogMapLt showing notable recall improvements from 0.50 in sharp to 0.63 in continuous. However, both slightly drop in precision in uncertain metrics, reflecting a cautious approach to confidence assignments.

Matcha and MDMapper adapt well to uncertain matches but face precision challenges. Matcha sees high recall improvement from 0.67 in sharp to 0.77 in continuous, though with a precision dip in uncertain metrics. MDMapper maintains stable recall across settings but loses precision, indicating confidence struggles with uncertain matches.

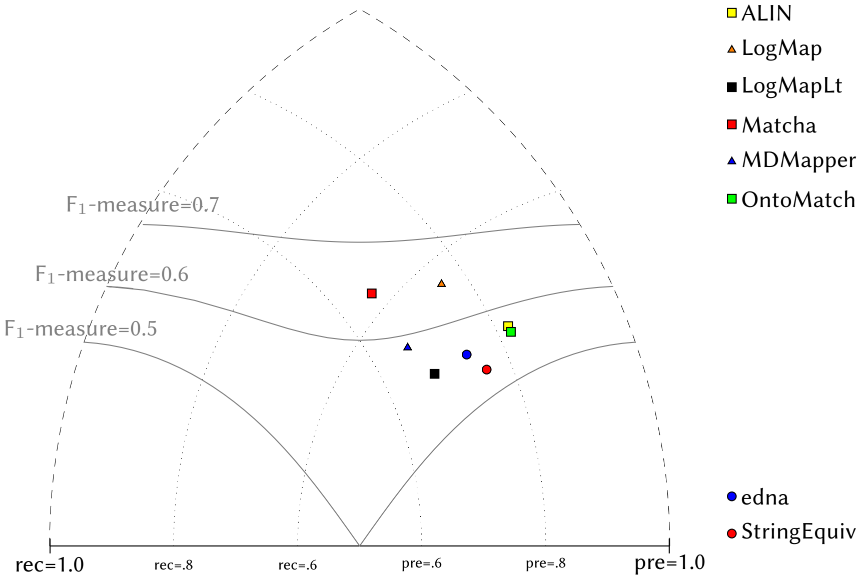


Figure 2: Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F$_1$-measure are depicted by areas bordered by corresponding lines F$_1$-measure=0.[5|6|7].

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

	Sharp			Discrete			Continuous		
System	Prec	F-ms	Rec	Prec	F-ms	Rec	Prec	F-ms	Rec
ALIN	0.88	0.61	0.47	0.88	0.70	0.59	0.87	0.71	0.60
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
Matcha	0.71	0.67	0.67	0.65	0.69	0.77	0.68	0.71	0.75
MDMapper	0.71	0.62	0.55	0.66	0.65	0.64	0.69	0.66	0.64
OntoMatch	0.88	0.60	0.46	0.88	0.69	0.57	0.88	0.70	0.59
TOMATO	0.61	0.51	0.44	0.61	0.58	0.56	0.61	0.59	0.56

TOMATO is the weakest performer, with limited recall and precision improvements, indicating difficulty capturing high-consensus matches confidently.

Overall, ALIN, LogMap, and OntoMatch excel with uncertain alignments, while TOMATO and MDMapper struggle, highlighting the need for confidence in uncertain data evaluation.

4.4. Multifarm

This year, 4 systems have registered to participate in the Multifarm track: LogMap, LogMapLt, Matcha and MDMapper. The number of participating tools is similar with respect to the last 4 campaigns (4 in 2023, 5 in 2022, 6 in 2021, 6 in 2020, 5 in 2019). This year, we lost the participation of LSMatch Multilingual. But we received new participation from MDMapper. 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 Windows Server 2025 machine configured with 96GB 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.
LogMap	13	.72	.42	.32
LogMapLt	265	.24	.038	.02
Matcha	309	.21	.28	.44
MDMapper	493	.25	.04	.26

running under a Intel Xeon Silver 4114 @2.20Ghz CPU, Tesla P40 GPU. 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:

The results (Table 12) indicate notable differences in performance across the four systems (LogMap, LogMapLt, Matcha, and MDMapper) with regard to processing time, precision, F-measure, and recall. LogMap exhibits the shortest processing time (13 minutes) and achieves the highest precision (0.72), but its recall is relatively low (0.32), resulting in a moderate F-measure of 0.42. LogMapLt takes significantly longer (265 minutes) but shows much lower precision (0.24) and a minimal F-measure (0.038), along with a low recall (0.02). Matcha requires even more time (309 minutes) and has a relatively balanced performance, with a precision of 0.21, an F-measure of 0.28, and the highest recall among the systems (0.44). Finally, MDMapper has the longest runtime (493 minutes) with low precision (0.25), recall (0.26), and an F-measure of 0.04, indicating limited effectiveness despite the extended processing time. Overall, LogMap stands out for its efficiency and higher precision, while Matcha demonstrates better recall, albeit at a significant cost in processing time.

4.5. Complex Matching

Unfortunately, this track has not attracted many participants in the last years. This year only CANARD has registered to participate. As CANARD depends on instances, it has only run on the Populated Conference dataset. The system has been improved since its last participation in OAEI (2018), by adopting embeddings generated by LLM. Table 13 shows the results of CANARD 2024 together with a comparison to systems participating in the previous campaign (AMLC and Matcha-DL).

Matcher	Precision	Coverage
Matcha-DL	-	-
AMLC	0.230	0.260
CANARD 2018	0.212	0.471
CANARD 2024 (Stella-base IE 0.85)	**0.389**	0.623
CANARD 2024 (GritLM-7B ESQ)	0.359	**0.679**

Table 13
Precision in this table stands for classical precision and Coverage to classical - query F-measure coverage.

The results show that the integration of LLMs enhances the performance of CANARD, by increasing the precision and F-measure by up to 45% over the baseline (CANARD 2018). These results corroborate the effectiveness of such models in capturing semantic nuances. The configurations with Stella-base model on the Instance Embeddings (IE) component and GritLM-7B model on the Embeddings of SPARQL Query (ESQ) component of CANARD were the most effective.

4.6. Food

This is the third year of the track and three systems were registered: LogMap, LogMapLt and Matcha.

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

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

System	Corresp.	Precision	Recall	F1-measure	Time(s)
"equal" (=) relation					
LogMap	15	0.1333	0.0274	0.0454	20
LogMapLt	15	0.1333	0.0274	0.0454	7
Matcha	360	0.0611	0.3013	0.1016	47
"subclass" relation (<) relation					
LogMap	15	0	0	0	17
LogMapLt	15	0	0	0	7
Matcha	335	0	0	0	49

evaluated systems compute the alignment in less than a minute. LogMapLt stands out for its very fast calculation time of 7s to find "equal" (resp. "subclass" relation correspondences). Concerning "equal" (=) relation correspondences, LogMap and LogMapLt have better precision than Matcha. However, LogMap's recall is 20 (resp. 11 times) less than 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 15 and Figure 3 for both the Anatomy and Conference datasets.

Table 15
Interactive matching results for the Anatomy and Conference datasets.

Tool	Error	Prec.	Rec.	F-m.	Rec.+	Prec. oracle	Rec. oracle	F-m. oracle	Tot. Reqs.	Dist. Mapps	Pos. Prec.	Neg. Prec.
Anatomy Dataset												
ALIN	NI	0.983	0.726	0.835	0.438	–	–	–	–	–	–	–
	0.0	0.986	0.878	0.929	0.678	0.986	0.878	0.929	262	699	1.0	1.0
	0.1	0.953	0.863	0.905	0.655	0.986	0.876	0.928	235	626	0.79	0.957
	0.2	0.924	0.851	0.886	0.648	0.986	0.88	0.93	235	629	0.64	0.903
	0.3	0.904	0.836	0.869	0.62	0.986	0.878	0.929	221	590	0.53	0.838
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	0.967	0.831	0.894	0.567	0.97	0.802	0.878	388	1164	0.745	0.966
	0.2	0.951	0.822	0.881	0.602	0.951	0.761	0.846	388	1164	0.563	0.925
	0.3	0.937	0.816	0.873	0.538	0.926	0.726	0.814	388	1164	0.428	0.87
Conference Dataset												
ALIN	NI	0.874	0.456	0.599	–	–	–	–	–	–	–	–
	0.0	0.915	0.702	0.794	–	0.915	0.702	0.794	187	557	1.0	1.0
	0.1	0.753	0.664	0.705	–	0.926	0.724	0.812	181	537	0.562	0.984
	0.2	0.631	0.64	0.635	–	0.935	0.748	0.831	179	531	0.356	0.969
	0.3	0.539	0.612	0.573	–	0.942	0.767	0.846	176	522	0.236	0.944
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	0.597	0.702	–	0.859	0.578	0.691	82	246	0.71	0.973
	0.2	0.824	0.593	0.69	–	0.829	0.545	0.657	82	246	0.506	0.946
	0.3	0.797	0.585	0.675	–	0.808	0.518	0.631	82	246	0.37	0.908

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

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

- The performance of the system: Precision (Prec.), Recall (Rec.), and F-measure (F-m.) with respect to the fixed reference alignment, as well as Recall+ (Rec.+) for the Anatomy task. To facilitate the assessment of the impact of user interactions, we also provide the performance results from the original tracks, without interaction (line with Error NI).
- To ascertain the impact of the oracle errors, we provide the performance of the system with respect to the oracle (i.e., the reference alignment as modified by the errors introduced by the oracle: Precision oracle (Prec. oracle), Recall oracle (Rec. oracle) and F-measure oracle (F-m. oracle). For a perfect oracle, these values match the actual performance of the system.
- Total requests (Tot Reqs.) represents the number of distinct user interactions with the tool, where each interaction can contain one to three conflicting correspondences, that could be 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).
- 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 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 [59]: 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

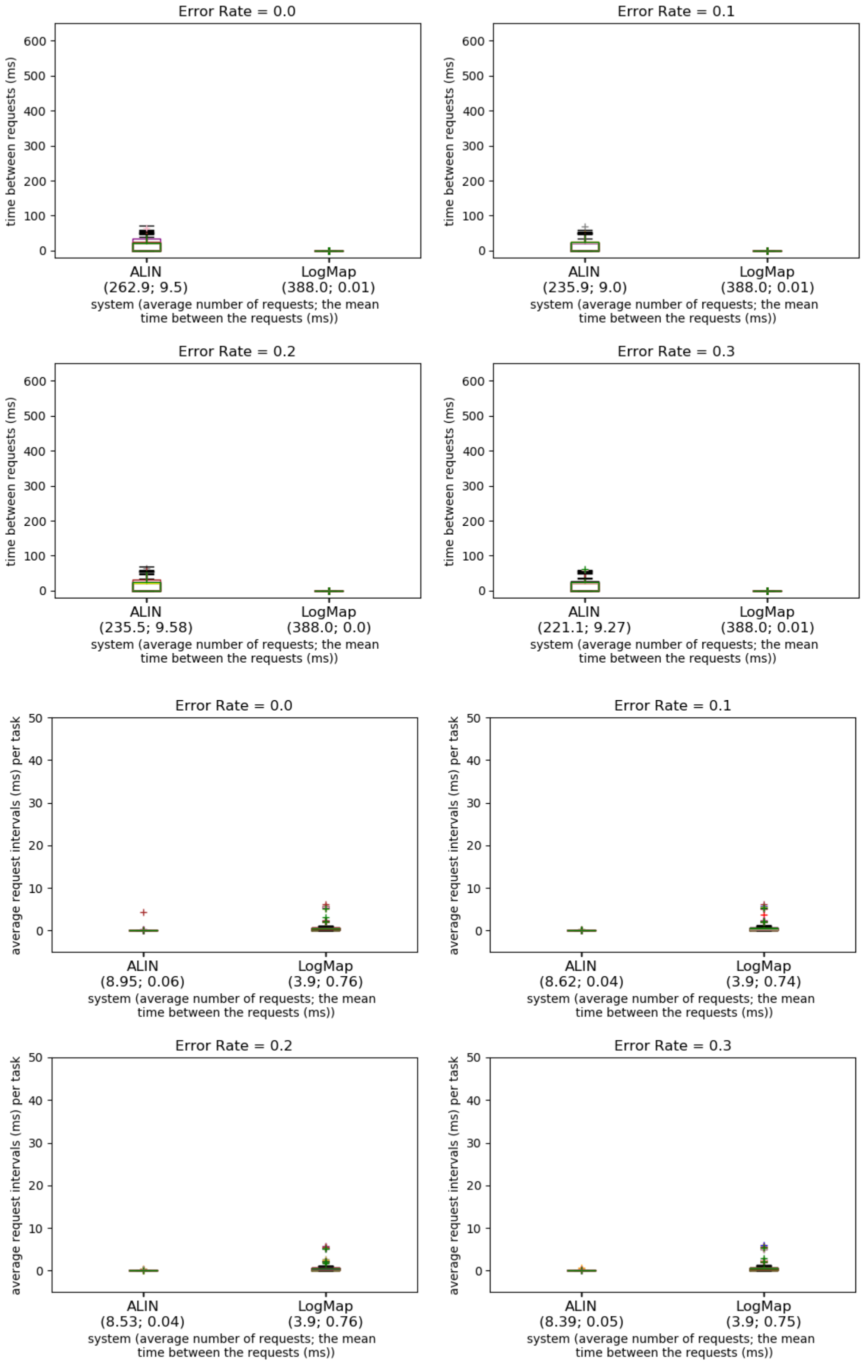


Figure 3: Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

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. See Table 16 for an overview of the equivalence matching results. For the full results, please refer to the OAEI 2024 Bio-ML website[42].

Briefly, we have the following participants for equivalence matching: *(i)* machine learning-based systems including BERTMap, BERTMapLt [57], BioGITOM, BioSTransMatch, HybridOM and Matcha [60, 61]; and *(ii)* traditional systems including LogMap, LogMapBio, LogMapLt [39].

In equivalence matching, the top-performing systems varied across tasks. BioGITOM achieved the highest F1 score in 3 out of 5 semi-supervised tasks, while HybridOM led in the remaining two. For unsupervised tasks, LogMapBio and HybridOM each attained the best F1 score in 2 out of 5 tasks, with BERTMap excelling in the last one. Notably, BERTMap also achieved the best ranking scores on most tasks, although some systems did not provide ranking results for equivalence matching. In subsumption matching, no new systems participated this year.

In summary the 2024 edition saw the introduction of three new machine learning-based systems. While some participants from previous years did not resubmit their systems, the increased number of machine learning-based participants aligns with Bio-ML's original mission. Meanwhile, LogMap variants remained the only symbolic systems in the competition.

4.9. Biodiversity and Ecology

This year, four matching systems (LogMap, LogMapLt, LogMapKG, and Matcha) 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 17.

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. Digital Humanities

Matcha, LogMap, LogMap Bio and LogMap KG found alignments. TOMATO and LogMap lite were running without code errors, but resulted in empty alignments. ALIN and MDMapper had code exceptions when executing. The same happened when trying to run OntoMatch, which was standalone and not possible to run with MELT / SEALS.

When comparing the matching systems (see table 18), LogMap KG has the best averaged F1-score of 0.61. It is noteworthy that LogMap KG's performance is more stable across tracks compared to the second best, Matcha. The latter performed very well in some test cases, but poor in others.

When we examine the F1-scores averaged over all matchers (see table 19), they range from 0.24 to 0.77. This indicates that while the matchers perform fairly well on some test cases, there is considerable room for improvement on others.

Looking at execution times (see table 20), they are all in the same range, between 13s and 21s to run the full track. The only exception is LogMap lite with over 20 min but still results in an empty alignment.

In general, less than half of the evaluated matchers, and only one matcher that is not based on LogMap, can find alignments. Most of the systems resulted in errors, which aligns with our findings in our related OM-paper [62] where only five out of 17 systems could find alignments. This makes it

[42] https://krr-oxford.github.io/OAEI-Bio-ML/2024/index.html

Table 16
Results for the Bio-ML track, systems that do not use training maps in the semi-supervised setting are marked with an asterisk (*).

Task	Method	Unsupervised		Semi-supervised	
		F-score	MRR	F-score	MRR
OMIM-ORDO	BERTMap	0.646	**0.88**	0.617	**0.891**
	BERTMapLt *	0.623	0.766	0.615	0.766
	BioGITOM	-	-	**0.853**	-
	BioSTransMatch	0.407	0.741	0.432	0.737
	HybridOM*	0.685	0.849	0.645	0.849
	LogMap*	0.593	-	0.589	-
	LogMapBio*	**0.715**	-	0.703	-
	LogMapLt*	0.397	-	0.407	-
	Matcha*	0.617	0.815	0.602	0.815
NCIT-DOID	BERTMap	0.883	**0.959**	0.856	**0.96**
	BERTMapLt*	0.839	0.89	0.825	0.89
	BioGITOM	-	-	**0.913**	-
	BioSTransMatch	0.735	0.9	0.719	0.906
	HybridOM*	**0.918**	0.952	0.904	0.952
	LogMap*	0.779	-	0.767	-
	LogMapBio*	0.908	-	0.879	-
	LogMapLt*	0.725	-	0.723	-
	Matcha*	0.814	0.902	0.792	0.902
SNOMED-FMA	BERTMap	**0.79**	0.944	0.792	**0.965**
	BERTMapLt*	0.785	0.892	0.787	0.892
	BioGITOM	-	-	**0.923**	-
	BioSTransMatch	0.192	0.633	0.464	0.855
	HybridOM*	**0.79**	0.907	0.772	0.907
	LogMap*	0.526	-	0.511	-
	LogMapBio*	0.68	-	0.66	-
	LogMapLt*	0.696	-	0.693	-
	Matcha*	0.641	**0.95**	0.63	0.95
SNOMED-NCIT (Pharm)	BERTMap	0.73	**0.969**	0.796	**0.971**
	BERTMapLt*	0.724	0.849	0.718	0.849
	BioGITOM	-	-	0.827	-
	BioSTransMatch	0.69	0.943	0.852	0.957
	HybridOM*	**0.902**	0.964	**0.885**	0.964
	LogMap*	0.746	-	0.738	-
	LogMapBio*	0.737	-	0.724	-
	LogMapLt*	0.748	-	0.743	-
	Matcha*	0.752	0.936	0.746	0.936
SNOMED-NCIT (Neoplas)	BERTMap	0.643	**0.954**	0.65	**0.962**
	BERTMapLt*	0.752	0.891	0.729	0.891
	BioGITOM	-	-	-	-
	BioSTransMatch	0.402	0.846	0.65	0.855
	HybridOM*	0.755	0.911	**0.732**	0.911
	LogMap*	0.701	-	0.683	-
	LogMapBio*	**0.771**	-	0.729	-
	LogMapLt*	0.67	-	0.662	-
	Matcha*	0.665	0.889	0.642	0.889

evident that most matching systems cannot handle SKOS even though SKOS is widely used in research across different fields. This issue was already noted in the early library tracks [63] but has yet to be addressed.

Table 17
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
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
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
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
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
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
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
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
Matcha	00:00:44	493	0.724	1.0	0.840

4.11. Archaeology multilingual

Since this track is composed of datasets of the digital humanities track, the matching systems that found alignments resp. resulted in errors are identical in both tracks, therefore see section 4.10 for more information.

Comparing the matching systems (see table 21), LogMap Bio, and LogMap KG perform best with an averaged F1-score of 0.26. This is in particular surprising for LogMap Bio because it was originally developed / tuned for another domain.

When looking at the F1-scores averaged over all matchers (see table 22), they range from 0.00 (finding no alignments) to 0.59. Only the language combinations English-English and German-German are on the upper end, while all the others are at or below 0.24. It is particularly interesting that for the language combination French-Italian, not a single matching system was able to find alignments. The results suggest that most systems struggle when dealing with different languages. The fact that German and English both belong to the West Germanic languages might be advantageous. The romance languages pose a bigger challenge that the systems cannot solve in large parts.

Table 18
Matching system performance for the digital humanities (dh) track. The numbers are rounded to two decimal places. The best performing matcher of each test case is highlighted.

Test Case	Precision				Recall				F1-score			
	Log-Map	Log-Map Bio	Log-Map KG	Mat-cha	Log-Map	Log-Map Bio	Log-Map KG	Mat-cha	Log-Map	Log-Map Bio	Log-Map KG	Mat-cha
defc-pactols	0.33	0.20	0.90	**1.00**	**1.00**	0.20	0.90	0.90	0.50	0.20	0.90	**0.95**
idai-pactols	0.35	**0.40**	**0.40**	0.31	**1.00**	0.71	0.71	0.24	**0.52**	0.51	0.51	0.27
ironage...-pactols	0.31	0.40	0.40	**0.67**	**0.80**	**0.80**	**0.80**	**0.80**	0.44	0.53	0.53	**0.73**
pactols-parthenos	0.42	0.71	0.71	**0.83**	**0.92**	0.83	0.83	0.83	0.58	0.77	0.77	**0.83**
idai-parthenos	0.70	**1.00**	**1.00**	0.00	**0.27**	0.17	0.17	0.00	**0.39**	0.30	0.30	0.00
oeai-parthenos	0.51	**1.00**	**1.00**	0.90	**0.89**	0.68	0.68	0.74	0.65	**0.81**	**0.81**	**0.81**
dha-unesco	0.25	**0.50**	**0.50**	0.08	**0.90**	0.40	0.40	0.60	0.39	**0.44**	**0.44**	0.14
tadirah-unesco	0.22	0.00	**0.53**	0.48	**0.80**	0.00	0.67	0.67	0.35	0.00	**0.59**	0.56
Average over all tracks	0.39	0.53	**0.68**	0.53	**0.82**	0.47	0.64	0.60	0.48	0.45	**0.61**	0.54

Table 19
Averaged evaluation metrics over all matchers for each test case of the digital humanities (dh) track.

Test case	Precision	Recall	F1-Score
arch1_defc-pactols	0.61	0.75	0.64
arch2_idai-pactols	0.36	0.66	0.45
arch3_ironagedanube-pactols	0.44	0.80	0.56
arch4_pactols-parthenos	0.67	**0.85**	0.74
cult1_idai-parthenos	0.68	0.15	0.24
cult2_oeai-parthenos	**0.85**	0.75	**0.77**
dhcs1_dha-unesco	0.33	0.58	0.36
dhcs2_tadirah-unesco	0.31	0.53	0.37
Average over all tracks	0.53	0.63	0.52

Table 20
Total runtime for all test cases of the digital humanities (dh) track.

Test case	total runtime (hh:mm:ss)
LogMap	00:00:13
LogMap Bio	00:00:15
LogMap KG	00:00:11
LogMap lite	00:22:16
Matcha	00:00:15
TOMATO	00:00:21

The execution times (see table 23) are below half a minute for the whole track, except Matcha (2h31min) and LogMap lite (55 min), while the latter only resulted in empty alignments.

It can clearly be seen that handling languages other than English needs to be addressed by matching systems. This is particularly important for making matching systems useful for domains like the Digital Humanities where research objects are in multiple languages and the research itself is frequently conducted in the local language of the respective research institution.

4.12. Circular Economy

Three systems have been registered for the first year of the Circular Economy track: LogMap, LogMapLt, and Matcha. We conducted experiments by executing each system in its standard setting, and we compared precision, F-measure, and recall. We used the MELT platform to execute our evaluations for all systems.

Table 24 shows the results for precision, F-measure, recall and the size of the alignments for the optimal threshold. Regarding the recall, Matcha achieved the best score. LogMap and Matcha provide the

Table 21
Matching system performance for the archaeology multilingual track. The numbers are rounded to two decimal places. The best performing matcher in each test case is highlighted.

Test Case	Precision				Recall				F1-score			
	Log-Map	Log-Map Bio	Log-Map KG	Mat-cha	Log-Map	Log-Map Bio	Log-Map KG	Mat-cha	Log-Map	Log-Map Bio	Log-Map KG	Mat-cha
idai-pactols_de-de	0.85	0.91	0.91	**1.00**	**0.65**	0.59	0.59	0.12	**0.73**	0.71	0.71	0.21
idai-pactols_de-en	0.25	**0.33**	**0.33**	0.02	**0.06**	**0.06**	**0.06**	**0.06**	**0.10**	**0.10**	**0.10**	0.03
idai-pactols_de-fr	**0.40**	**0.40**	**0.40**	0.04	0.12	0.12	0.12	**0.18**	**0.18**	**0.18**	**0.18**	0.07
idai-pactols_de-it	**0.50**	**0.50**	**0.50**	0.00	**0.12**	**0.12**	**0.12**	0.00	**0.19**	**0.19**	**0.19**	0.00
idai-pactols_en-en	0.27	0.60	0.60	**0.75**	**0.67**	0.50	0.50	0.50	0.38	0.55	0.55	**0.60**
idai-pactols_en-fr	0.13	**1.00**	**1.00**	0.03	0.17	0.17	0.17	**0.33**	0.14	**0.29**	**0.29**	0.05
idai-pactols_en-it	0.50	**1.00**	**1.00**	0.00	**0.17**	**0.17**	**0.17**	0.00	0.25	**0.29**	**0.29**	0.00
idai-pactols_fr-fr	0.09	0.13	0.13	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	0.13	0.17	0.17	**0.25**
idai-pactols_fr-it	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
idai-pactols_it-it	0.17	0.10	0.10	**0.30**	**0.75**	0.25	0.25	**0.75**	0.27	0.14	0.14	**0.43**
Average over all tracks	0.31	**0.50**	**0.50**	0.24	**0.29**	0.22	0.22	0.22	0.24	**0.26**	**0.26**	0.16

Table 22
Averaged evaluation metrics over all matchers for each test case of the archaeology multilingual track.

Test case	Precision	Recall	F1-Score
idai-pactols_de-de	**0.92**	0.49	**0.59**
idai-pactols_de-en	0.23	0.06	0.08
idai-pactols_de-fr	0.31	0.13	0.15
idai-pactols_de-it	0.38	0.09	0.14
idai-pactols_en-en	0.55	**0.54**	0.52
idai-pactols_en-fr	0.54	0.21	0.19
idai-pactols_en-it	0.63	0.13	0.21
idai-pactols_fr-fr	0.15	0.25	0.18
idai-pactols_fr-it	0.00	0.00	0.00
idai-pactols_it-it	0.17	0.50	0.25
Average over all tracks	0.39	0.24	0.23

Table 23
Total runtime for all test cases of the archaeology multilingual track.

Test case	total runtime (hh:mm:ss)
LogMap	00:00:13
LogMap Bio	00:00:19
LogMap KG	00:00:12
LogMap lite	00:55:03
Matcha	02:31:50
TOMATO	00:00:27

correspondences with real-valued confidence. Therefore, we applied thresholding during the evaluation.

Table 24
The results for the circular economy track.

System	Size	Precision	F_1-measure	Recall
Matcha (0.9)	28	0.393	0.478	0.611
LogMapLt	29	0.379	0.468	0.611
LogMap (0.5)	23	0.391	0.439	0.5

The weights in LogMap's alignment range from 0.93 to 0.14 (with only one weight below 0.5). The mapping with the highest weight was a false positive, so was the mapping with the lowest weight. There were multiple matches with the second lower weight (0.5). These mappings were a mix of correct mappings and false positives. Based on these results, an optimal threshold for LogMap's results could be set to 0.5 (including) which is also the computed threshold with the highest F-measure.

In case of Matcha, the weights of its results range between 1 and 0.600378464. Mappings with the highest weights were both correct and false positives. The correct mapping with the lowest weight was weighted to 0.65293388. This weight is just a little higher than the lowest weight. However, most true positives have weights greater than 0.9. Based on these results, an optimal threshold for Matcha's results could be set to 0.9 which is also the computed threshold with the highest F-measure.

Additionally, we analysed the false positives - alignments discovered by the tools which were evaluated as incorrect. Looking at the results, it can be said that when the reason for discovering an alignment was the *same name*, all or at least most tools generated the mapping. LogMap and Matcha further generated mappings based on *similar strings.* All three systems generated mappings where the *same word* was present in the entities' names. Lastly, Matcha produced 2 mappings where the reason is not obvious. As a possibly interesting observation, there were no false positives found which would be generated based on synonyms in entities' names. More information is provided at the results web page.

The first evaluation within the track shows that matching circular economy relevant ontologies remains a challenging task for tools (F1-measure lower than 0.48). Based on false positives analysis, it turns out that mere string matching could be misleading, and the meaning of entities should be better considered.

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[43] (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.

Table 25 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.

The resulting alignments are available for download [44]. This year's best overall system is still the baseline using the alternative labels (0.84 F-measure). The highest recall is again achieved by Matcha (0.84). Detailed results for each test case can be found on the OAEI results page of the track[45].

Property matches are still not created by all systems. LogMap, Matcha, and MDMapper 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`.

When it comes to class matches, Matcha is the overall best system with an F-measure of 0.87 (much better than the provided baseline).

For further analysis of the results, we also provide an online dashboard[46] generated with MELT[64]. In this dashboard, the results can be inspected on a correspondence level. Due to the large amount of these correspondences, it can take some time to load the full website.

Regarding runtime, Matcha (38:48:16) and LogMapLt (64:48:07) were the slowest systems. Besides the baselines (which need around 12 minutes for all test cases) LogMap (00:56:43) is the fastest system.

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

[44] http://oaei.ontologymatching.org/2024/results/knowledgegraph/knowledgegraph-alignments.zip

[45] http://oaei.ontologymatching.org/2024/results/knowledgegraph/index.html

[46] http://oaei.ontologymatching.org/2024/results/knowledgegraph/knowledge_graph_dashboard.html

Table 25
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)
Matcha	38:48:16	5	23.8	0.97 (0.97)	0.87 (0.87)	0.80 (0.80)
MDMapper	02:28:53	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
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)
Matcha	38:48:16	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
MDMapper	02:28:53	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)
Matcha	38:48:16	5	249510.0	0.55 (0.55)	0.63 (0.63)	0.86 (0.86)
MDMapper	02:28:53	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)
Matcha	38:48:16	5	249533.8	0.55 (0.55)	0.63 (0.63)	0.84 (0.84)
MDMapper	02:28:53	5	24.6	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

4.14. Pharmacogenomics

For this second year of the Pharmacogenomics track, only LogMap registered with its different versions: LogMap, LogMap-Bio, LogMap-Lite, and LogMap-KG. The evaluation was performed using the MELT framework.

None of these versions were successful in producing alignments between reified n-ary tuples. We identify the main reason as the absence of labels for n-ary tuples, since providing labels allows the LogMap versions to produce alignments. However, when labels are present, altering neighborhoods does not impact the produced alignments, showing that only labels are taken into account by the different versions of the LogMap system. Recall that n-ary tuples are reifed as abstract entities because RDF does not allow n-ary relations. Hence, labels of such reified entities are seldom present in general, but their neighbors play a crucial role in their identity. This makes us conclude that submitted systems are not adequate to the task of matching pharmacogenomic knowledge as they appear to rely only on labels and disregard neighbors. It is also noteworthy that, without adding labels for n-ary tuples, some versions of LogMap output alignments between other entities (*e.g.*, components of pharmacogenomic tuples) but not between the n-ary tuples themselves. Such alignments are valid but sometimes trivial (*e.g.*, between entities in the two ontologies to match that actually share the same URI) and out of the scope of the Pharmacogenomics track.

5. Conclusions and Lessons Learned

As in previous campaigns, 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 in the last years, 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.

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 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 results of the **Digital Humanities** track clearly show that SKOS vocabularies are not well-supported by most matching systems. Regarding the matching systems that were able to find alignments, there is still room to improve, especially when the results are used for more complex alignment and mapping tasks. To further improve this track, it is planned to include more subdomains of the digital humanities.

The **Archaeology multilingual** track leads to the conclusion that different languages within SKOS vocabularies are not adequately supported, especially when coming to the family of Romance languages. In future track versions, it is aimed for including additional ancient languages like Latin or Ancient Greek.

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 7 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, only one 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.

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.

For the second year of the **Pharmacogenomics** track, participation was limited with only four versions of a single system registered, namely LogMap. None of these versions were successful in producing alignments between reified n-ary tuples, which, according to our investigation, is due to the absence of labels for the tuples of our dataset. These results highlight again the interest in considering domain-specific problems, bringing additional challenges to the field of ontology matching (here, different types of alignments between individuals, structure-based matching). Given the inadequacy of registered systems to produce valid alignments, such challenges are currently unaddressed and require to design new methods like [55, 65] or enrich existing ones. This ultimately motivates to propose again the track in the next editions of OAEI, hoping to attract new systems targeting this real-world matching scenario.

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 [25] (since 2023 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 [66]. 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.

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).

Eva Blomqvist, Patrick Lambrix, Huanyu Li, Ondřej Zamazal and Jana Vataščinová have been supported by the European Union's Horizon Europe research and innovation programme under grant agreement no. 101058682 (Onto-DESIDE).

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.

The work of Felix Kraus was funded by the research program "Engineering Digital Futures" of the Helmholtz Association of German Research Centers, and the Helmholtz Metadata Collaboration Platform (HMC).

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, A. Coulet, J. Cufi, H. Dong, O. Fallatah, D. Faria, I. Fundulaki, S. Hertling, Y. He, I. Horrocks, M. Huschka, L. Ibanescu, S. Jain, E. Jiménez-Ruiz, N. Karam, P. Lambrix, H. Li, Y. Li, P. Monnin, E. Nasr, H. Paulheim, C. Pesquita, T. Saveta, P. Shvaiko, G. Sousa, C. Trojahn, J. Vatascinova, M. Wu, B. Yaman, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2023, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn (Eds.), Proceedings of the 18th International Workshop on Ontology Matching (OM 2023) co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023, volume 3591 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 97–139. URL: https://ceur-ws.org/Vol-3591/oaei23_paper0.pdf.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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. doi:10.1186/s13326-017-0166-5.

[27] 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.

[28] 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.

[29] É. Thiéblin, O. Haemmerlé, C. Trojahn, Automatic evaluation of complex alignments: An instance-based approach, Semantic Web 12 (2021) 767–787. URL: https://doi.org/10.3233/SW-210437. doi:10.

3233/SW-210437.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

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

[39] 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.

[40] 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).

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] S. M. Winslow, G. Schneider, R. Bleier, C. Steiner, C. Pollin, G. Vogeler, Ontologies in the Digital Repository: Metadata Integration, Knowledge Management and Ontology-Driven Applications, in: A. Barton, S. Seppälä, D. Porello, R. Ferrario, E. M. Sanfilippo, M. Nicolosi Asmundo (Eds.), Proceedings of the Joint Ontology Workshops 2019, volume 2518 of *CEUR Workshop Proceedings*, CEUR, Graz, Austria, 2019.

[47] J. Euzenat, P. Shvaiko, Ontology Matching, second edition ed., Springer, Berlin, Heidelberg, 2013. doi:10.1007/978-3-642-38721-0.

[48] E. Blomqvist, H. Li, R. Keskisärkkä, M. Lindecrantz, M. A. N. Pour, Y. Li, P. Lambrix, Cross-domain Modelling–A Network of Core Ontologies for the Circular Economy, in: Proceedings of the 14th Workshop on Ontology Design and Patterns (WOP 2023) co-located with the 22nd International Semantic Web Conference (ISWC 2023), CEUR Workshop Proceedings, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3636/paper1.pdf.

[49] C. Bicchielli, N. Biancone, F. Ferri, P. Grifoni, BiOnto: An Ontology for Sustainable Bioeconomy and Bioproducts , Sustainability 13 (2021) 4265. doi:10.3390/su13084265.

[50] H. Li, M. Abd Nikooie Pour, Y. Li, M. Lindecrantz, E. Blomqvist, P. Lambrix, A Survey of General Ontologies for the Cross-Industry Domain of Circular Economy, in: Companion Proc. of the ACM Web Conference 2023, ACM, 2023. doi:10.1145/3543873.3587613.

[51] H. Li, E. Blomqvist, P. Lambrix, Initial and Experimental Ontology Alignment Results in the Circular Economy Domain, in: Proceedings of the 2nd International Workshop on Knowledge Graphs for Sustainability (KG4S2024), CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3753/short1.pdf.

[52] S. Hertling, H. Paulheim, Dbkwik: extracting and integrating knowledge from thousands of wikis, Knowledge and Information Systems (2019).

[53] S. Hertling, H. Paulheim, Dbkwik: A consolidated knowledge graph from thousands of wikis, in: Proceedings of the International Conference on Big Knowledge, 2018.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] D. Faria, M. C. Silva, P. Cotovio, P. Eugénio, C. Pesquita, Matcha and matcha-dl results for oaei

2022., in: OM@ ISWC, 2022, pp. 197–201.

[61] D. Faria, M. C. Silva, P. Cotovio, L. Ferraz, L. Balbi, C. Pesquita, Results for matcha and matcha-dl in oaei 2023., in: OM@ ISWC, 2023, pp. 164–169.

[62] F. Kraus, N. Blumenröhr, G. Götzelmann, D. Tonne, A. Streit, A Gold Standard Benchmark Dataset for Digital Humanities, in: Proceedings of the 19th International Workshop on Ontology Matching, CEUR Workshop Proceedings, Baltimore, USA, in press.

[63] 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: P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt (Eds.), Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008), volume 431 of *CEUR Workshop Proceedings*, CEUR-WS.org, Karlsruhe, Germany, 2008.

[64] 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.

[65] 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.

[66] 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: Proceedings of the 18th International Workshop on Ontology Matching (OM 2023) co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023, volume 3591 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3591/om2023_STpaper3.pdf.

Linköping, Jena, Lisboa, Heraklion, Mannheim, Montpellier, Oslo, London, Berlin, Trento, Toulouse,
Prague, Manhattan, Dublin, Grenoble, Oxford, Karlsruhe
December 2024