Results of the Ontology Alignment Evaluation Initiative 2025

Mina Abd Nikooie Pour[1,2], Eva Blomqvist[1], Pedro Giesteira Cotovio[3,4], Adrien Coulet[5,6], Lucas Ferraz[4], Sven Hertling[7], Sarika Jain[8], Ernesto Jiménez-Ruiz[3], Felix Kraus[9], Patrick Lambrix[1,2], Huanyu Li[1], Ying Li[1,2], Xianhao Liu[10,11], Pierre Monnin[12], Heiko Paulheim[7], Catia Pesquita[4], Abhisek Sharma[8], Pavel Shvaiko[13], Marta Silva[4], Guilherme Sousa[14], Cassia Trojahn[15], Jana Vataščinová[16], Beyza Yaman[17], Ondřej Zamazal[16] and Lu Zhou[18]

[1]*Department of Computer and Information Science, Linköping University, Linköping, Sweden*
[2]*Swedish e-Science Research Centre, Linköping, Sweden*
[3]*City St George's, University of London, UK*
[4]*LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[5]*Inria Paris, France*
[6]*Centre de Recherche des Cordeliers, Inserm, Université Paris Cité, Sorbonne Université, France*
[7]*Data and Web Science Group, University of Mannheim, Germany*
[8]*National Institute of Technology Kurukshetra, Haryana, India*
[9]*Karlsruhe Institute of Technology, Karlsruhe, Germany*
[10]*Stibo Systems, Denmark*
[11]*Technical University of Denmark, Denmark*
[12]*Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France*
[13]*Trentino Digitale SpA, Trento, Italy*
[14]*Institut de Recherche en Informatique de Toulouse, France*
[15]*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France*
[16]*Prague University of Economics and Business, Czech Republic*
[17]*ADAPT Centre, Trinity College Dublin*
[18]*Apple Inc., USA*

Abstract

The Ontology Alignment Evaluation Initiative (OAEI) aims at comparing ontology matching systems on precisely defined test cases. These test cases can be based on ontologies of different levels of complexity and use different evaluation modalities. The OAEI 2025 campaign offered 12 tracks and was attended by 20 participants. This paper is an overall presentation of that campaign.

1. Introduction

The Ontology Alignment Evaluation Initiative[1] (OAEI) is a coordinated international initiative, which organizes the evaluation of ontology matching systems [1, 2], and 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

OM 2025: The 20th International Workshop on Ontology Matching collocated with the 24th International Semantic Web Conference (ISWC 2025), November 2nd, 2025, Nara, Japan

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

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, 23], which this year took place in Nara, Japan.[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. During 2017 to 2023, a novel evaluation environment called HOBBIT [24] 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 [25] 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 2025 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 two alternative platforms: the SEALS client, or the MELT framework. Both 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 (see Section 3.5), 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 **MELT framework**[4] [25] was introduced in 2019 and is under active development. It allows the development, evaluation, and packaging of matching systems for evaluation interfaces like SEALS or HOBBIT. It further enables developers to use Python or any other programming language in their matching systems, which, beforehand, had been a hurdle for OAEI participants. The evaluation client[5] allows organizers to evaluate packaged systems whereby multiple submission formats are supported (SEALS packages or matchers implemented as Web services). Starting from OAEI 2023, the MELT framework also supports the SSSOM [26] 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, two submission formats were allowed: (1) SEALS package, and (2) MELT. With the increasing usage of other programming languages than Java and increasing hardware re-

[2] http://om.ontologymatching.org/2025
[3] http://www.seals-project.eu
[4] https://github.com/dwslab/melt
[5] https://dwslab.github.io/melt/matcher-evaluation/client

quirements for matching systems, since 2021, the MELT Web interface has been introduced to address this issue. It mainly consists of a technology-independent HTTP interface[6] 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.

Since 2024, we also allowed submission of 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 2025 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}, 2025. 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}, 2025, at which date participants had to submit the (near) final versions of their systems.

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 12 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.
- 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.

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

Table 1
Tracks in OAEI 2025.

test	formalism	relations	confidence	modalities	language	SEALS	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, HI, TR		√
complex	OWL	=	[0 1]	open+blind	EN, ES		
interactive	OWL	=, <=	[0 1]	open	EN	√	
bio-ML	OWL	=	[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		√
beyond equivalence	OWL	=	[0 1]	open	EN		√
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[7] (3304 classes) and the anatomy of the mouse[8] (2744 classes). The evaluation is based on a manually curated reference alignment. This dataset has been used since 2007 with some improvements over the years [27].

Systems are evaluated with the standard parameters of precision, recall, F-measure. Additionally, recall+ is computed by excluding trivial correspondences (i.e., correspondences that have the same normalized label). Alignments are also checked for coherence using the Pellet reasoner. The evaluation was carried out on a machine with a 5 core CPU @ 1.80 GHz with 16GB allocated RAM, using the MELT framework. For some systems, the SEALS client has been used. However, the evaluation parameters were computed *a posteriori*, after removing from the alignments produced by the systems, correspondences expressing relations other than equivalence, as well as trivial correspondences in the oboInOwl namespace (e.g., oboInOwl#Synonym = oboInOwl#Synonym). The results obtained with the SEALS client vary in some cases by 0.5% compared to the results presented in Section 4.2.

3.2. Conference

The conference track 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 [28].

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.

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

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 [29] 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 12 languages: Arabic (ar), Chinese (cn), Czech (cz), Dutch (nl), French (fr), German (de), Italian (it), Portuguese (pt), Russian (ru), Spanish (es), Hindi(hi), and Turkish (tr). 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 Windows machine configured with 16GB 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* $\sqcap$ $\exists o_2$*:hasDecision.*o_2*:Acceptance*.

The track ran with eight datasets: Conference and Populated Conference, Hydrography, GeoLink and Populated GeoLink, Populated Enslaved, and Biomedical, a new dataset introduced this year for complex multi-ontology matching.

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** dataset 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.[9] The **Hydrography** dataset is composed of four tasks, where four source ontologies (Hydro3, HydrOntology_native, HydrOntology_translated, and Cree) are aligned with a single target ontology, the Surface Water Ontology (SWO). The **GeoLink** dataset includes a single matching task between the GeoLink Base Ontology (GBO) and the GeoLink Modular Ontology (GMO) [30]. The **Populated GeoLink** dataset is based on the previous one and adds instance data from

[9] https://framagit.org/IRIT_UT2J/conference-dataset-population

seven data repositories in the GeoLink project [31]. The **Populated Enslaved** dataset is composed of two ontologies, the Enslaved Ontology and the Enslaved Wikidata Knowledge Graph, where the instance data originates from the Wikidata repository and the consensus was obtained from domain experts from several historian research institutions [32]. The **Biomedical** dataset is a new addition this year for the specific task of complex multi-ontology matching, and it is based on the existing logical definitions from three different biomedical ontologies: the Human Phenotype ontology (HP), the Mammalian Phenotype ontology (MP), and the Worm Phenotype ontology (WBP) [33]. Both HP and MP have the same set of target ontologies: Cell ontology (CL), Chemical Entities of Biological Interest (ChEBI), Gene Ontology (GO), Phenotype and Trait Ontology (PATO), and Uber Anatomy Ontology (UBERON). WBP uses ChEBI, GO, PATO, and the *C.elegans* Gross Anatomy Ontology (WBbt).

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. Three new evaluation strategies were debuted this year: Class evaluation [34], Graph Edit Distance [35], and Tree Edit Distance [36], which were run in the remaining tasks.

3.5. Interactive Matching

The interactive matching track aims to assess the performance of semi-automated matching systems by simulating user interaction [37, 38, 39]. 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 [40, 38].

The interactive matching track is based on the datasets from the Anatomy and Conference tracks, which have been previously described. It relies on the SEALS client's *Oracle* class to simulate user interactions. An interactive matching system can present a collection of correspondences simultaneously to the oracle, telling the system whether that correspondence is correct or not. If a system presents up to three correspondences together and each correspondence presented has a mapped entity (i.e., class or property) in common with at least one other correspondence presented, the oracle counts this as a single interaction, under the rationale that this corresponds to a scenario where a user is asked to choose between conflicting candidate correspondences. To simulate the possibility of user errors, the oracle can be set to reply with a given error probability (randomly, from a uniform distribution). We evaluated systems with four different error rates: 0.0 (perfect user), 0.1, 0.2, and 0.3. In addition to the standard evaluation parameters, we also compute the number of requests made by the system, the total number of distinct correspondences asked, the number of positive and negative answers from the oracle, the performance of the system according to the oracle (to assess the impact of the oracle errors on the system) and finally, the performance of the oracle itself (to assess how erroneous it was).

The evaluation was carried out on a server with 3.46 GHz (6 cores) and 8GB RAM allocated to the matching systems. For systems requiring more RAM, the evaluation was carried out on a computer with an AMD Ryzen 7 5700G 3.80 GHz CPU and 32GB RAM, with 10GB of max heap space allocated to Java. Each system was run ten times and the final result of a system for each error rate represents the average of these runs. For the Conference dataset with the ra1 alignment, precision and recall correspond to the micro-average over all ontology pairs, whereas the number of interactions is the total number of interactions for all the pairs.

3.6. Bio-ML

The Bio-ML track [41] incorporates *equivalence* ontology matching (OM) tasks for biomedical ontologies, with ground truth (equivalence) mappings extracted from Mondo [42] and UMLS [43] (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 [41], we make available five OM pairs with their information reported in Table 3. Each OM pair is accompanied with equivalence matching tasks; each matching task has two data split settings, i.e., *unsupervised* setting with no training mappings, and *semi-supervised*

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[10]	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[11]	358,222
	FMA	BioPortal, V4.14.0	104,523
	NCIT	BioPortal, V21.02d	163,842

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

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

setting with 30% ground truth mappings for training/validation. Since the 2023 edition, Bio-ML has added a *logical module enrichment* [44] 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 [41] 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.

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[12] [45] library.

3.7. 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 [2]. 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 second time. It includes eight test cases from archaeology, cultural history and DH / computer science. Each test case consists of two SKOS (using RDF/XML as 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 4 and Table 5.

[10]Created from OMIM texts by Mondo's pipeline tool avaiable at: https://github.com/monarch-initiative/omim.
[11]Created by the official snomed-owl-toolkit available at: https://github.com/IHTSDO/snomed-owl-toolkit.
[12]https://krr-oxford.github.io/DeepOnto/#/

Table 4
Controlled vocabularies used for the digital humanities (dh) track.

Resource	Field[13]	Version / Date	#concepts[14]	language (ISO 639)
DEFC Thesaurus[15]	Archaeology	-	∼800	de, en, la
PACTOLS thesaurus for archaeology[16]	Archaeology	- / 2021-05-18	∼60,000	ar, de, en, es, fr, it, nl
Iron-Age-Danube thesaurus[17]	Archaeology	1 / 2018-11-07	∼6900	de, en, hr, hu, sl
iDAI.world Thesaurus[18]	Arch. / cult. hist.	1.2 / 2022-02-10	∼290	de, en, es, fr, it
PARTHENOS Vocabularies[19]	Arch. / cult. hist.	- / 2019-05-07	∼4200	en
OeAI Thesaurus - Cultural Time Periods[20]	Cultural history	1.0.0 / 2022-11-23	∼400	de, en
DHA Taxonomy[21]	DH/CS	- / 2018-04-03	∼120	en
UNESCO[22]	DH/CS	- / 2024-06-03	∼4500	ar, en, fr, es, ru
TaDiRAH[23]	DH/CS	2.0.1 / 2021-07-22	∼170	de, en, es, fr, it, pt, sr

Table 5
Properties of the digital humanities (dh) track.

Domain	Source (#terms[24])	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

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.

3.8. Archaeology Multilingual

The archaeology multilingual track is based on an archaeology test case of the digital humanities track, see Section 3.7, with focus on evaluating matcher performance when dealing with different languages.

Like the DH track, this track participated for the second time. Each test case uses iDAI.world and PACTOLS (see Table 4 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 3.7.

[13]This is the field to which the CV was grouped within our dataset.
[14]This is the number of concepts in the primary language of the CV before any preprocessing steps.
[15]https://vocabs.dariah.eu/defc_thesaurus/en/
[16]https://isl.ics.forth.gr/bbt-federated-thesaurus/PACTOLS/en/
[17]https://vocabs.dariah.eu/iad_thesaurus/en/
[18]https://isl.ics.forth.gr/bbt-federated-thesaurus/DAI/en/
[19]https://vocabs.dariah.eu/parthenos_vocabularies/en/
[20]https://vocabs.acdh.oeaw.ac.at/oeai-cp/en/
[21]https://vocabs.dariah.eu/dha_taxonomy/en/
[22]https://vocabularies.unesco.org/browser/thesaurus/en/
[23]https://vocabs.dariah.eu/tadirah/en/
[24]The number of terms varies depending on the branch used for the respective domain.

3.9. 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. The circular economy track was proposed since 2024, and included 2 tasks in this year. In both tasks CE-specific ontology is matched to the Circular Economy Ontology Network (CEON) [47]. The first task is to match CEON to the Sustainable Bioeconomy and Bioproducts Ontology (BiOnto) [48]. The second task is to match CEON to materials domain ontology, MatOnto [49]. CEON (including 214 classes) from the Onto-DESIDE project,[25] aims to represent core concepts for the CE domain [47]. BiOnto (including 780 classes) from the BIOVOICES project,[26] focuses on establishing a shared and common terminology in the bioeconomy domain. MatOnto (848 classes) covers the materials domain. Materials are a central concept in circular value networks. In all, using BiOnto or MatOnto different stakeholders participating circular value networks can provide information according to ontologies [50].

The evaluation was 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, 52] and further validated by ontology engineers and CE domain experts from Onto-DESIDE project. The results is presented in Section 4.10.

3.10. Beyond Equivalence

This is the first time the Beyond Equivalence track is being organized. The goal of this track is to evaluate the ability of ontology matching systems to detect correspondences beyond simple equivalence. Specifically, systems are tasked with identifying five mutually disjoint relation types: *Equivalence* ($\equiv$), *Superclass_of* ($\leq$), *Subclass_of* ($\geq$), *Overlap* ($\simeq$), and *Disjointness* ($\perp$).

Benchmark data for the track is drawn from two main sources. First, a variety of industrial product classification schemes (e.g., GPC,[27] UNSPSC,[28] ETIM,[29], and eClass[30]), collectively referred to as "Product Classification Standards". Second, a diverse set of test-cases derived from the STROMA/TaSeR [53, 54] repository, consists of d 5 datasets: g1-web, g2–diseases, g3–text, g5–groceries, and g7–literature. Altogether, the benchmark suite comprises multiple datasets of varying size, structure, and semantic complexity; in each case the reference alignments are annotated with explicit relation types beyond simple equivalence. Table 6 presents the statistic information of the datasets.

Beyond Equivalence track encourages the development of more powerful and semantically-aware matching systems — in line with applications such as knowledge-graph merging, master data integration, and semantic search, where ontologies frequently differ in granularity, structure, or conceptual perspective.

3.11. Knowledge Graph

The Knowledge Graph track was run for the fifth 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[31] in the course of the DBkWik project [55, 56]. 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

[25] https://ontodeside.eu
[26] https://www.biovoices.eu
[27] https://gpc-browser.gs1.org/
[28] https://www.undp.org/unspsc
[29] https://www.etim-international.com/
[30] https://eclass.eu/en/
[31] https://www.wikia.com/

Table 6
Datasets for the OAEI 2025 Beyond Equivalence track, with counts of classes, properties, and relation-type annotations in the reference alignments (equivalence $\equiv$, superclass $\geq$, subclass $\leq$, overlap $\simeq$).

Dataset	# S:Class	# T:Class	# S:Property	# T:Property	$\equiv$	$\geq$	$\leq$	$\simeq$
Product Classification Standards								
GPC–UNSPSC	1,151	30,707	839	3	255	1,377	250	17,738
GPC–UNSPSC+	5,342	30,707	839	3	381	3,306	1,293	23,600
ETIM–eClass	2,877	5,565	10,746	8,932	2,371	380	636	291
eClass–UNSPSC	7,349	19,600	4,009	3	329	1,533	351	55,573
eClass–GPC	3,459	1,210	2,027	2,064	251	431	1,199	11,014
STROMA/TaSeR								
g1–web	728	1,132	0	0	175	26	29	—
g2–diseases	1,109	5,146	0	0	316	11	27	—
g3–text	335	260	0	0	70	267	425	—
g5–groceries	60	335	0	0	29	113	14	—
g7–literature	41	155	0	0	12	52	18	—

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

from the wiki page itself. Due to the fact that not all interwiki links on a page represent the same concept, a few restrictions were made: 1) only links in sections with a header containing "link" are used, 2) all links are removed where the source page links to more than one concept in another wiki (ensures the alignments are functional), 3) multiple links which point to 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.[32]

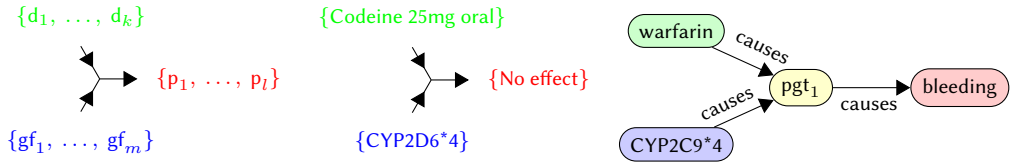
The alignments were evaluated based on precision, recall, and F-measure for classes, properties, and instances (each in isolation). The partial gold standard contained 1:1 correspondences, and we further assume that in each knowledge graph, only one representation of the concept exists. This means that 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.[33]

3.12. Pharmacogenomics

In 2025, the Pharmacogenomics track was run for the third 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.

[32] https://github.com/dwslab/melt/tree/master/examples/kgEvalCli
[33] 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.

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 [57]:

- 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 on 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 the same neighbors, thus represent the same two "pharmacogenomics relationships", and thus should be detected as identical.

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 `hemorrhagee` 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 [57, 58] 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[34] [59]. 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.

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

Table 8
Participants and the status of their submissions.

System	Agent-OM	ALIN	BioGITOM	BioSTransMatch	CMatch	DogMa	DRAL-OA	GraphMatcher	LogMap	LogMap-Bio	LogMapLt	LogMapKG	LogMapLLM	LSMatch	LSMatch-Multilingual	Matcha	MDMaper	NeSyMatch	OWL2Vec4OA	TIM	Total=20
anatomy	●	●	○	○	○	○	●	○	●	●	●	●	●	●	○	●	●	○	○	○	11
conference	●	●	○	○	○	○	○	○	●	○	●	○	○	●	○	●	●	○	○	○	7
multifarm	○	○	○	○	○	○	○	○	●	○	●	○	○	○	●	●	○	○	○	○	4
complex	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	●	○	○	○	○	2
interactive	○	●	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	2
bio-ml	◐	○	●	●	○	○	○	○	●	●	●	○	●	○	○	●	○	○	◐	○	9
knowledge graph	○	○	○	○	○	●	○	○	○	○	◐	●	○	●	○	◐	○	○	○	●	6
dh	●	○	○	○	○	○	○	○	●	●	○	●	○	○	○	●	○	○	○	●	6
arch-multiling	●	○	○	○	○	○	○	○	●	●	○	●	○	○	○	●	○	○	○	●	6
circular economy	●	○	○	○	○	○	○	○	●	○	●	○	○	○	○	●	○	○	○	○	4
pharmacogenomics	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	0
beyond equivalence	○	○	○	○	○	○	○	○	●	●	○	●	○	○	○	●	●	○	○	○	5
total	6	3	1	1	1	1	1	0	9	5	6	5	2	3	1	10	3	0	1	3	

4. Results and Discussion

4.1. Participation

Following an initial period of growth, the number of OAEI participants has remained approximately constant since 2012. This year we count with 20 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 [60] and BERTSubs [61]) 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, but without an OAEI system submission. Some matching systems participated with different variants (e.g., LogMap and LSMatch), whereas others were evaluated with different configurations, as requested by developers (see test case sections for details). The following sections summarize the results for each track.

4.2. Anatomy

The results for the Anatomy track are shown in Table 9. Among the 11 systems participating in the Anatomy track, 10 achieved an F-measure higher than the StringEquiv baseline. Three systems were first-time participants (i.e., Agent-OM, DRAL-OA, LogMapLLM) 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 ALIN which increased in F-measure (from 0.851 to 0.912), recall (from 0.75 to 0.884), recall+ (from 0.489 to 0.7), size (from 1156 to 1423), and decreased in precision (from 0.984 to 0.942). LogMap-Bio increased in size (from 1549 to 1561), recall (from 0.908 to 0.911), recall+ (from 0.757 to 0.766), and decreased in precision (from 0.888 to 0.885). MDMapper increased in size (from 1441 to 1483), recall+ (from 0.703 to 0.707), and decreased in precision (from 0.926 to 0.899), F-measure (from 0.903 to 0.889), recall (from 0.881 to 0.879). In terms of runtime, 5 out of 11 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 (Agent-OM, ALIN, and LogMapLLM) which is at least as good as the best systems in OAEI 2007-2010. Like in previous years, there is no significant correlation between the quality of the generated alignment and the run time. Four systems produced coherent alignments (i.e., LogMap, LogMap-Bio, LogMapKG and LogMapLLM).

Table 9
Anatomy track 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	47	1485	0.951	0.941	0.931	0.82	-
Agent-OM	-	1396	0.959	0.92	0.883	0.698	-
ALIN	1004	1423	0.942	0.912	0.884	0.7	-
LogMapLLM	500	1324	0.964	0.899	0.842	0.588	+
LogMap-Bio	1750	1561	0.885	0.898	0.911	0.766	+
MDMapper	124	1483	0.899	0.889	0.879	0.707	-
LogMap	8	1402	0.917	0.881	0.848	0.602	+
LogMapKG	9	1402	0.917	0.881	0.848	0.602	+
LogMapLt	2	1147	0.962	0.828	0.728	0.288	-
DRAL-OA	877	1509	0.83	0.828	0.827	0.56	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-
LSMatch	16	1009	0.952	0.761	0.634	0.037	-

4.3. Conference

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

With regard to two baselines we can group tools according to system's position: there are five matchers above (or equal to) edna baseline (ALIN, LogMap, Matcha, Agent-OM, and MDMapper), and two matchers below edna baseline but above StringEquiv baseline (LogMapLt, and LSMatch). Two matchers (MDMapper, and LSMatch) 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 systems, five (Agent-OM, ALIN, LogMapLt, LSMatch, and MDMapper) use 1.0 to all correspondences. The remaining 2 systems (LogMap, Matcha) have a wide variation of confidence values. Agent-OM internally applied a threshold of 0.9 before submission of alignments.

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.
ALIN	0.62	0.63	0.65	0.67	0.68	7	111	107
LogMap	0.76	0.71	0.64	0.59	0.56	0	21	0
Matcha	0.77	0.71	0.63	0.57	0.53	9	90	115
Agent-OM	0.64	0.63	0.61	0.6	0.59	8	101	136
MDMapper	0.69	0.64	0.58	0.53	0.5	3	81	39
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
LSMatch	0.83	0.69	0.55	0.46	0.41	0	2	0
StringEquiv	0.76	0.65	0.53	0.45	0.41			

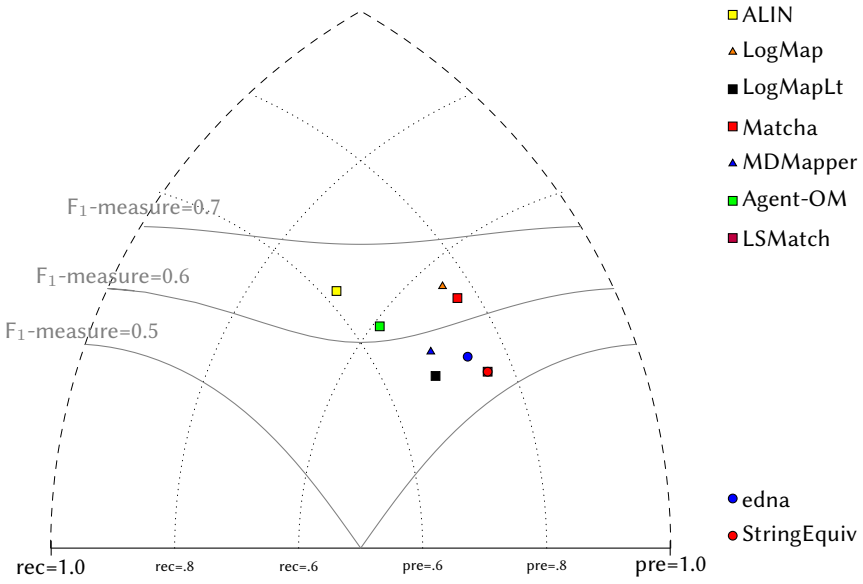


Figure 2: Precision/recall triangular graph for the conference test case. Dotted lines depict level of precision/recall while values of F_1-measure are depicted by areas bordered by corresponding lines F_1-measure=0.[5|6|7].

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
Agent-OM	0.67	0.63	0.60	0.64	0.68	0.72	0.61	0.65	0.71
ALIN	0.66	0.68	0.70	0.61	0.70	0.81	0.57	0.67	0.79
LogMap	0.81	0.68	0.58	0.81	0.70	0.62	0.80	0.66	0.58
LogMapLt	0.73	0.59	0.50	0.73	0.67	0.62	0.72	0.67	0.63
LSMatch	0.88	0.57	0.42	0.88	0.66	0.53	0.88	0.67	0.54
Matcha	0.85	0.68	0.56	0.60	0.67	0.77	0.64	0.69	0.75
MDMapper	0.75	0.62	0.53	0.72	0.68	0.64	0.72	0.67	0.63

Systems using fixed confidences show clear gains from the sharp to the uncertain evaluations. As in 2024, the discrete metric particularly benefits these systems by downweighting low-consensus matches. ALIN, LogMapLt, and MDMapper maintain balanced precision and recall, while LSMatch achieves the highest precision overall (0.88) and converts it into strong F-measure improvements. Agent-OM also performs competitively, showing robust recall despite lower precision.

Systems assigning graded confidences continue to perform best under the continuous metric. Their ability to model uncertainty enables stable precision and greater recall than in the sharp setting, confirming the advantage of calibrated confidence outputs.

Overall, the 2025 results reinforce two trends observed in 2024: (1) recall gains under uncertain evaluation are broader and more consistent, narrowing performance gaps among systems, and (2) both fixed- and graded-confidence approaches can benefit from uncertainty, provided alignments reflect the majority consensus.

4.4. Multifarm

This year, 4 systems have registered to participate in the Multifarm track: LogMap, LogMapLt, Matcha, and LSMatch-Multilingual. The number of participating tools is similar with respect to the last 4

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	6.7	.87	.18	.10
LogMapLt	16.9	.84	.008	.004
LSMatch-Multilingual	36	.79	.44	.30
Matcha	408	.26	.26	.25

campaigns (4 in 2024, 4 in 2023, 5 in 2022, 6 in 2021). This year, we welcome back LSMatch-Multilingual. The reader can refer to the OAEI papers for a detailed description of the strategies adopted by each system.

The Multifarm evaluation results based on the blind dataset are presented in Table 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 machine configured with 16GB of RAM running under a Intel Core i7-9750H @2.60Ghz CPU. 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 LSMatch-Multilingual) with regard to processing time, precision, F-measure, and recall. LogMap exhibits the shortest processing time (6.7 minutes) and achieves the highest precision (0.87), but its recall is relatively low (0.10), resulting in a moderate F-measure of (0.18). LogMapLt takes longer (16.9 minutes) but shows lower precision (0.84) and a minimal F-measure (0.008), along with a low recall (0.004). Matcha requires even more time (408 minutes) and has a relatively balanced performance, with a precision of 0.26, an F-measure of 0.26, and the highest recall among the systems (0.25). Finally, LSMatch-Multilingual has the runtime of (36 minutes) with precision (0.79), best recall (0.30), and hence best F-measure of 0.44, indicating limited effectiveness despite the extended processing time. Overall, LogMap stands out for its efficiency and higher precision, while LSMatch-Multilingual demonstrates better recall, and F-measure.

4.5. Complex Matching

Regarding *Conference dataset* (non-populated) sub-track, we had only one participant: Matcha. Matcha delivered *simple equivalences*, *complex correspondences*, and *subsumptions*. Within this track only complex correspondences have been evaluated. With regard to complex correspondences no TPs were identified. All complex correspondences contained intersections (of classes and also of classes and properties). While the intersection of classes is an important construct, EDOAL does not allow for directly intersecting a class and a property. Instead, class restrictions should be applied, such as *AttributeDomainRestriction*.

For the remaining tasks there were two participating systems: CMatch and Matcha. Additionally, the alignments from previous participating systems were added as placeholder baselines: AMLC, AROA, and CANARD (from the 2020 OAEI edition). The results for the evaluation metrics of Graph Edit Distance (GED) and Tree Edit Distance (TED) are shown in Table 13. As not all systems participated in all datasets it is difficult to draw general conclusions. CMatch has increased precision but low recall, leading to an F-measure that is lower than other systems in most cases. CANARD appears to be the more consistent choice across both metrics, with Matcha achieving far lower results with the TED metric but comparable ones with the GED metric. AMLC achieves the worse results out of all systems,

Table 13
Complex track results per matcher and per dataset for two of the new evaluation strategies.

	Graph Edit Distance			Tree Edit Distance		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
——— cmt-conference ———						
AMLC	0.557	0.080	0.139	0.000	0.000	0.000
CANARD	0.284	0.464	0.352	0.183	0.566	0.277
CMatch	1.000	0.029	0.056	0.000	0.000	0.000
Matcha	0.526	0.365	0.431	0.000	0.000	0.000
——— cmt-ekaw ———						
AMLC	0.255	0.060	0.097	0.000	0.000	0.000
CANARD	0.275	0.330	0.300	0.218	0.468	0.297
CMatch	0.615	0.181	0.280	0.667	0.176	0.279
Matcha	0.529	0.327	0.404	0.000	0.000	0.000
——— conference-ekaw ———						
AMLC	0.374	0.108	0.167	0.000	0.000	0.000
CANARD	0.252	0.256	0.254	0.171	0.506	0.256
Matcha	0.587	0.595	0.591	0.000	0.000	0.000
——— cree-swo ———						
CMatch	0.000	0.000	0.000	0.286	0.013	0.025
Matcha	0.500	0.064	0.113	0.077	0.011	0.019
——— hydro3-swo ———						
CMatch	0.319	0.154	0.208	0.561	0.228	0.325
Matcha	0.645	0.507	0.568	0.000	0.000	0.000
——— hydrOntology_native-swo ———						
Matcha	0.372	0.062	0.106	0.079	0.004	0.008
——— hydrOntology_translated-swo ———						
CMatch	0.307	0.118	0.170	0.553	0.033	0.062
Matcha	0.421	0.303	0.353	0.569	0.202	0.299
——— gbo-gmo ———						
AMLC	0.000	0.000	0.000	0.013	0.004	0.006
AROA	0.472	0.492	0.482	0.706	0.267	0.388
CMatch	0.644	0.176	0.276	0.773	0.056	0.104
Matcha	0.287	0.467	0.355	0.003	0.002	0.002
——— popgbo-popgmo ———						
CANARD	0.481	0.453	0.467	0.448	0.182	0.258
Matcha	0.245	0.391	0.301	0.002	0.004	0.003
——— enslaved-wikidata ———						
CANARD	0.049	0.188	0.078	0.218	0.166	0.189
Matcha	0.048	0.688	0.091	0.001	0.002	0.001
——— hp ———						
Matcha	0.465	0.465	0.465	-	-	-
——— mp ———						
Matcha	0.529	0.529	0.529	-	-	-
——— wbp ———						
Matcha	0.554	0.554	0.554	-	-	-

and it is interesting to note that AROA outperforms all other systems in the single task it participated in. As the submission was done as alignments, conclusions cannot be drawn on the computational and runtime costs required by the systems.

4.6. Interactive matching

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

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

Table 14
Interactive matching results for the Anatomy and Conference datasets.

Tool	Error	Prec.	Rec.	F-m.	Rec.+	Prec. oracle	Rec. oracle	F-m. oracle	Tot. Reqs.	Dist. Mapps	Pos. Prec.	Neg. Prec.
Anatomy Dataset												
ALIN	NI	0.941	0.887	0.913	0.71	–	–	–	–	–	–	–
	0.0	0.986	0.883	0.932	0.691	0.986	0.883	0.932	278	746	1.0	1.0
	0.1	0.953	0.868	0.909	0.669	0.986	0.883	0.932	253	675	0.796	0.954
	0.2	0.924	0.852	0.887	0.644	0.986	0.881	0.931	238	638	0.64	0.903
	0.3	0.895	0.837	0.865	0.621	0.987	0.879	0.93	230	617	0.504	0.842
LogMap	NI	0.916	0.846	0.88	0.593	–	–	–	–	–	–	–
	0.0	0.988	0.846	0.912	0.595	0.988	0.846	0.912	388	1164	1.0	1.0
	0.1	0.967	0.829	0.893	0.564	0.971	0.804	0.88	388	1164	0.751	0.963
	0.2	0.949	0.822	0.881	0.549	0.952	0.764	0.848	388	1164	0.573	0.93
	0.3	0.936	0.818	0.873	0.544	0.93	0.725	0.814	388	1164	0.433	0.88
Conference Dataset												
ALIN	NI	0.647	0.685	0.665	–	–	–	–	–	–	–	–
	0.0	0.903	0.703	0.79	–	0.903	0.703	0.79	187	554	1.0	1.0
	0.1	0.744	0.676	0.708	–	0.917	0.732	0.814	185	547	0.56	0.989
	0.2	0.631	0.64	0.635	–	0.926	0.748	0.828	180	535	0.355	0.967
	0.3	0.537	0.618	0.574	–	0.934	0.767	0.842	176	523	0.236	0.95
LogMap	NI	0.818	0.59	0.686	–	–	–	–	–	–	–	–
	0.0	0.886	0.61	0.723	–	0.886	0.61	0.723	82	246	1.0	1.0
	0.1	0.845	0.597	0.7	–	0.86	0.58	0.693	82	246	0.698	0.978
	0.2	0.814	0.588	0.683	–	0.828	0.544	0.656	82	246	0.481	0.939
	0.3	0.798	0.593	0.68	–	0.816	0.525	0.639	82	246	0.385	0.926

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

- 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).

Figure 3 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 uses user interactions exclusively in the post-matching steps to filter their candidate

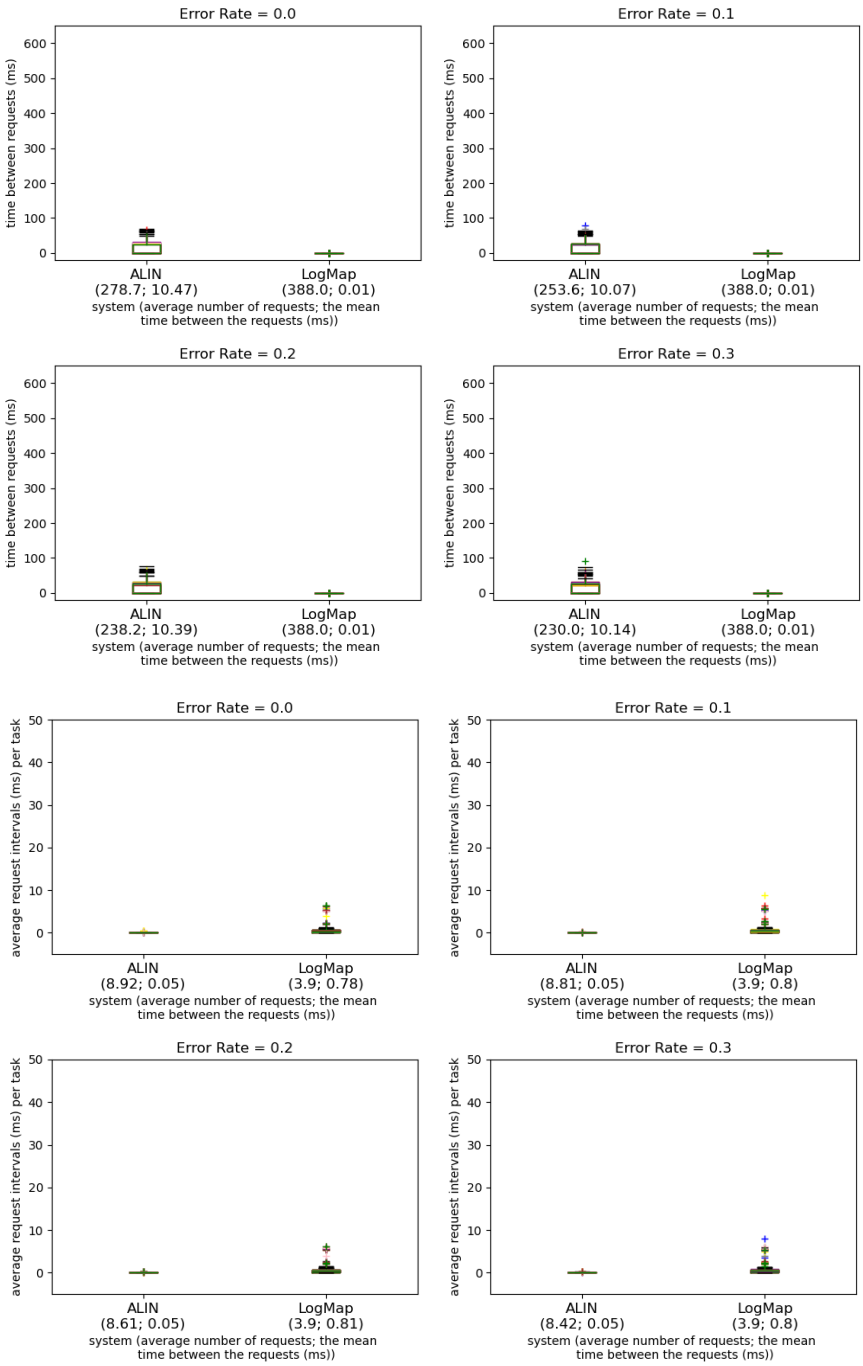


Figure 3: Time intervals between requests to the user/oracle for the Anatomy (top 4 plots) and Conference (bottom 4 plots) datasets. Whiskers: Q1-1,5IQR, Q3+1,5IQR, IQR=Q3-Q1. The labels under the system names show the average number of requests and the mean time between the requests for the ten runs.

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. The improvement of ALIN is mainly 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 [62]: Shneiderman and Seow take different approaches to categorize the response times taking a task-centered view and a user-centered view respectively. According to task complexity, Shneiderman defines response time in four categories: typing, mouse movement (50-150 ms), simple frequent tasks (1 s), common tasks (2-4 s) and complex tasks (8-12 s). While Seow's definition of response time is based on the user expectations towards the execution of a task: instantaneous (100-200 ms), immediate (0.5-1 s), continuous (2-5 s), captive (7-10 s). Ontology alignment is a cognitively demanding task and can fall into the third or fourth categories in both models. In this regard the response times (request intervals as we call them above) observed in all datasets fall into the tolerable and acceptable response times, and even into the first categories, in both models. The request intervals for LogMap and ALIN stay at a few milliseconds for most datasets. It could be the case, however, that a user would not be able to take advantage of these low response times because the task complexity may result in higher user response time (i.e., the time the user needs to respond to the system after the system is ready).

4.7. Bio-ML

Our results comprise five tables, where each table corresponds to a specific ontology matching (OM) pair and includes results for both the unsupervised and semi-supervised settings. An overview of the results is presented in Table 15. Full results are available on the OAEI 2025 Bio-ML website.[35]

Briefly, the participating systems are: *(i)* machine learning-based approaches, including Agent-OM [63], BERTMap, BERTMapLt [60], BioGITOM [64], BioSTransMatch, LogMap-LLM, Logmap+OWL2Vec4OA, Matcha [65, 66], and OWL2Vec4OA; and *(ii)* traditional symbolic systems, namely LogMap, LogMap-Bio, and LogMapLt [44].

The top-performing systems varied across tasks. BioGITOM achieved the highest F1 score in 3 out of 5 semi-supervised tasks, with BERTMap and LogMap-LLM each leading in one. For unsupervised tasks, LogMap-Bio and LogMap-LLM attained the best F1 score in 2 tasks each, with BERTMap leading the remaining one. Notably, BERTMap obtained the best ranking scores in all but one task, where Matcha took the top spot, although some systems did not provide ranking results.

In summary, the 2025 edition introduced four new machine learning-based systems. Although some participants from previous years did not resubmit their tools, the growing number of learning-based systems is consistent with Bio-ML's original mission. Meanwhile, LogMap and its variants remained the only symbolic systems participating in the campaign.

4.8. Digital Humanities

Among the submitted systems, Agent-OM, LogMap, LogMap-Bio, LogMapKG, Matcha and TIM successfully found alignments. LSMatch executed without runtime errors but generated empty alignments. ALIN and MDMapper encountered code exceptions during execution. Since Agent-OM can not be used within the MELT framework, it could not be executed or verified by the organizers. Instead, its alignments were provided directly by the system developers.

When comparing system performance (see Table 16), Matcha achieved the highest average F1-score of 0.64, an improvement of 0.05 over the best-performing system in the previous OAEI edition.

[35] https://liseda-lab.github.io/OAEI-Bio-ML/2025/index.html

Table 15
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
NCIT-DOID	Agent-OM*	0.753	-	0.742	-
	BERTMap	0.883	**0.959**	0.856	**0.96**
	BERTMapLt*	0.839	0.89	0.825	0.89
	BioGITOM	0.755	0.913	**0.918**	0.890
	BioSTransMatch	0.649	0.856	0.585	0.856
	LogMap*	0.867	-	0.885	-
	LogMap-Bio*	**0.908**	-	0.879	-
	LogMapLt*	0.725	-	0.723	-
	LogMap-LLM*	0.907	-	0.892	-
	Logmap+OWL2Vec4OA*	-	0.873	-	0.873
	Matcha*	0.814	0.902	0.792	0.902
	OWL2Vec4OA*	-	0.879	-	0.879
OMIM-ORDO	Agent-OM*	0.375	-	0.369	-
	BERTMap	0.646	**0.88**	0.617	**0.891**
	BERTMapLt*	0.623	0.766	0.615	0.766
	BioGITOM	0.640	0.834	**0.787**	0.834
	BioSTransMatch	0.309	0.693	0.251	0.693
	LogMap*	0.589	-	0.593	-
	LogMap-Bio*	**0.715**	-	0.703	-
	LogMap-LLM*	0.626	-	0.626	-
	LogMapLt*	0.397	-	0.407	-
	Logmap+OWL2Vec4OA*	-	0.692	-	0.692
	Matcha*	0.617	0.815	0.602	0.815
	OWL2Vec4OA*	-	0.707	-	0.707
SNOMED-FMA	BERTMap	**0.79**	0.944	**0.792**	**0.965**
	BERTMapLt*	0.785	0.892	0.787	0.892
	BioGITOM	0.646	0.909	0.787	0.909
	BioSTransMatch	0.250	0.633	0.196	0.633
	LogMap*	0.651	-	0.667	-
	LogMap-Bio*	0.68	-	0.66	-
	LogMap-LLM*	0.682	-	0.671	-
	LogMapLt*	0.696	-	0.693	-
	Matcha*	0.641	**0.95**	0.63	0.95
SNOMED-NCIT (Pharm)	Agent-OM*	0.446	-	0.428	-
	BERTMap	0.73	**0.969**	0.796	**0.971**
	BERTMapLt*	0.724	0.849	0.718	0.849
	BioGITOM	0.648	0.913	**0.786**	0.913
	BioSTransMatch	0.431	0.908	0.355	0.908
	LogMap*	0.745	-	0.755	-
	LogMap-Bio*	0.737	-	0.724	-
	LogMap-LLM*	**0.760**	-	0.753	-
	LogMapLt*	0.748	-	0.743	-
	Logmap+OWL2Vec4OA*	-	0.852	-	0.852
	Matcha*	0.752	0.936	0.746	0.936
	OWL2Vec4OA*	-	0.864	-	0.864
SNOMED-NCIT (Neoplas)	Agent-OM*	0.211	-	0.206	-
	BERTMap	0.643	**0.954**	0.65	**0.962**
	BERTMapLt*	0.752	0.891	0.729	0.891
	BioGITOM	0.617	0.929	0.745	0.929
	BioSTransMatch	0.295	0.779	0.237	0.779
	LogMap*	0.736	-	0.774	-
	LogMap-Bio*	0.771	-	0.729	-
	LogMap-LLM*	**0.782**	-	**0.754**	-
	LogMapLt*	0.670	-	0.662	-
	Logmap+OWL2Vec4OA*	-	0.808	-	0.808
	Matcha*	0.665	0.889	0.642	0.889
	OWL2Vec4OA*	-	0.828	-	0.828

Considering the average F1-scores across all matchers (see Table 17), values range from 0.31 to 0.61. This variation indicates that while several systems perform reasonably well on certain test cases, there remains substantial potential for improvement on others.

Regarding runtime performance (see Table 18), most systems completed the track in under 20s. Matcha was running about 20 times longer. LogMapKG offers the best balance between runtime and result quality. Agent-OM did not provide runtimes due to dependency on external APIs.

Overall, the number of systems capable of generating alignments increased by two compared to the previous year. Nevertheless, several systems still failed due to runtime errors or unsupported features. This observation is consistent with the findings reported in our previous OM study [67], in which only five out of seventeen systems were able to produce valid alignments. These results highlight that many ontology matching systems continue to face challenges when processing SKOS vocabularies. On a more positive note, both newly participating systems this year were able to successfully handle SKOS data.

Table 16
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					
	Agent-OM	LogMap	LogMap Bio	LogMap KG	Matcha	TIM	Agent-OM	LogMap	LogMap Bio	LogMap KG	Matcha	TIM	Agent-OM	LogMap	LogMap Bio	LogMap KG	Matcha	TIM
defc-pactols	**1.00**	0.33	0.20	0.90	**1.00**	0.13	0.20	**1.00**	0.20	0.90	0.90	0.60	0.33	0.50	0.20	0.90	**0.95**	0.21
idai-pactols	**1.00**	0.35	0.40	0.40	0.45	0.04	0.12	**1.00**	0.71	0.71	**1.00**	0.12	0.21	0.52	0.51	0.51	**0.63**	0.06
ironage...-pactols	**1.00**	0.31	0.40	0.40	0.31	0.04	0.20	**0.80**	**0.80**	**0.80**	0.24	0.40	0.33	0.44	**0.53**	**0.53**	0.27	0.08
pactols-parthenos	**1.00**	0.42	0.71	0.71	0.80	0.18	0.50	**0.92**	0.83	0.83	0.23	0.67	0.67	0.58	**0.77**	**0.77**	0.36	0.29
idai-parthenos	0.50	0.70	**1.00**	**1.00**	0.67	0.00	0.38	0.27	0.17	0.17	**0.80**	0.00	0.43	0.39	0.30	0.30	**0.73**	0.00
oeai-parthenos	0.91	0.51	**1.00**	**1.00**	0.90	0.00	0.43	**0.89**	0.68	0.68	0.74	0.00	0.58	0.65	**0.81**	**0.81**	**0.81**	0.00
dha-unesco	0.67	0.25	0.50	0.50	**0.83**	0.02	0.40	**0.90**	0.40	0.40	0.83	0.20	0.50	0.39	0.44	0.44	**0.83**	0.04
tadirah-unesco	**1.00**	0.22	0.00	0.53	0.36	0.00	0.27	0.80	0.00	0.67	**0.93**	0.00	0.42	0.35	0.00	**0.59**	0.52	0.00
Average over all tracks	**0.88**	0.39	0.53	0.68	0.67	0.05	0.31	**0.82**	0.47	0.65	0.71	0.25	0.43	0.48	0.45	0.61	**0.64**	0.09

Table 17
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	0.59	0.63	0.52
arch2_idai-pactols	0.44	0.61	0.41
arch3_ironagedanube-pactols	0.41	0.54	0.36
arch4_pactols-parthenos	0.64	**0.66**	0.57
cult1_idai-parthenos	0.65	0.30	0.36
cult2_oeai-parthenos	**0.72**	0.57	**0.61**
dhcs1_dha-unesco	0.46	0.52	0.44
dhcs2_tadirah-unesco	0.35	0.45	0.31
Average over all tracks	0.48	0.48	0.40

4.9. Archaeology Multilingual

Since this track builds directly on datasets from the digital humanities track, the set of systems that executed successfully is identical to those reported in Section 4.8.

Comparing the matching systems (see Table 19), Agent-OM achieved the highest average F1-score of 0.33, outperforming last year's best system by 0.07. TIM failed to produce alignments for almost all test cases.

Examining the F1-scores averaged across all matchers (Table 20), values range from 0.04 to 0.47. As expected, English–English and German–German combinations are handled most effectively, while test cases involving only non-English languages remain particularly difficult. Agent-OM shows promising results even for these test cases and is the first system to generate non-empty alignments for the French–Italian test case.

Table 18
Total runtime for all test cases of the digital humanities (dh) track.

Test case	total runtime (hh:mm:ss)
Agent-OM	not provided
LogMap	00:00:14
LogMap-Bio	00:00:17
LogMapKG	00:00:13
LogMapLt	00:24:52
LSMatch	00:00:17
LSMatch Multilingual	00:00:18
Matcha	00:05:25
TIM	00:00:07

Table 19
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					
	Agent-OM	LogMap	LogMap Bio	LogMap KG	Matcha	TIM	Agent-OM	LogMap	LogMap Bio	LogMap KG	Matcha	TIM	Agent-OM	LogMap	LogMap Bio	LogMap KG	Matcha	TIM
idai-pactols_de-de	0.75	0.85	0.91	0.91	**1.00**	0.00	0.35	**0.65**	0.59	0.59	0.12	0.00	0.48	**0.73**	0.71	0.71	0.21	0.00
idai-pactols_de-en	**0.50**	0.25	0.33	0.33	0.17	0.00	**0.29**	0.06	0.06	0.06	0.06	0.00	**0.37**	0.10	0.10	0.10	0.09	0.00
idai-pactols_de-fr	**0.67**	0.40	0.40	0.40	0.33	0.00	**0.12**	**0.12**	**0.12**	**0.12**	**0.12**	0.00	**0.20**	0.18	0.18	0.18	0.17	0.00
idai-pactols_de-it	0.33	**0.50**	**0.50**	**0.50**	0.40	0.00	0.06	**0.12**	**0.12**	**0.12**	**0.12**	0.00	0.10	**0.19**	**0.19**	**0.19**	0.18	0.00
idai-pactols_en-en	0.60	0.27	0.60	0.60	**0.75**	0.00	0.50	**0.67**	0.50	0.50	0.50	0.00	0.55	0.38	0.55	0.55	**0.60**	0.00
idai-pactols_en-fr	0.67	0.13	**1.00**	**1.00**	0.33	0.00	**0.33**	0.17	0.17	0.17	0.17	0.00	**0.44**	0.14	0.29	0.29	0.22	0.00
idai-pactols_en-it	0.33	0.50	**1.00**	**1.00**	0.50	0.00	**0.33**	0.17	0.17	0.17	0.17	0.00	**0.33**	0.25	0.29	0.29	0.25	0.00
idai-pactols_fr-fr	**0.25**	0.09	0.13	0.13	**0.25**	0.02	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	0.13	0.17	0.17	**0.25**	0.03
idai-pactols_fr-it	**0.20**	0.00	0.00	0.00	0.00	0.00	**0.25**	0.00	0.00	0.00	0.00	0.00	**0.22**	0.00	0.00	0.00	0.00	0.00
idai-pactols_it-it	0.22	0.17	0.10	0.10	**0.30**	0.02	0.50	**0.75**	0.25	0.25	**0.75**	0.25	0.31	0.27	0.14	0.14	**0.43**	0.04
Average over all tracks	0.45	0.31	**0.50**	**0.50**	0.40	0.00	**0.30**	0.29	0.22	0.22	0.23	0.05	**0.33**	0.24	0.26	0.26	0.24	0.01

Table 20
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	**0.74**	0.38	**0.47**
idai-pactols_de-en	0.26	0.09	0.13
idai-pactols_de-fr	0.37	0.10	0.15
idai-pactols_de-it	0.37	0.09	0.14
idai-pactols_en-en	0.47	**0.45**	0.44
idai-pactols_en-fr	0.52	0.17	0.23
idai-pactols_en-it	0.56	0.17	0.24
idai-pactols_fr-fr	0.15	0.25	0.17
idai-pactols_fr-it	0.03	0.04	0.04
idai-pactols_it-it	0.15	0.46	0.22
Average over all tracks	0.33	0.20	0.21

Execution times for most systems (see Table 21) are below half a minute for the whole track, except for Matcha and LogMapLt, with the latter producing only empty alignments.

The results clearly show that handling languages other than English remains an open challenge for ontology matching systems. Compared to last year, however, there is now one system (Agent-OM) that explicitly addresses this issue and achieves encouraging results across several multilingual test cases. Progress in this direction is essential, particularly for domains such as the Digital Humanities, where research data are multilingual and research by the scholars is frequently conducted in the local language of the respective institution.

4.10. Circular Economy

Four systems have been registered for the Circular Economy track: Agent-OM, LogMap, LogMapLt, and Matcha. We conducted experiments by executing each system in its standard setting, and we compared

Table 21
Total runtime for all test cases of the archaeology multilingual track.

Test case	total runtime (hh:mm:ss)
Agent-OM	not provided
LogMap	00:00:14
LogMap-Bio	00:00:18
LogMapKG	00:00:13
LogMapLt	01:02:30
LSMatch	00:00:22
LSMatch Multilingual	00:00:23
Matcha	00:10:42
TIM	00:00:10

Table 22
The results for the circular economy track.

System	Size	Precision	F_1-measure	Recall
LogMap	82	0.5	0.646	0.911
Agent-OM	81	0.469	0.603	0.844
Matcha	85 (0.9)	0.459	0.6	0.867
LogMapLt	88	0.42	0.556	0.822

precision, F-measure, and recall. We used the MELT platform to execute our evaluations for all systems. The only exception was Agent-OM which provided its own alignments (using commercial LLM).

Table 22 shows the results for precision, F-measure, recall and the size of the alignments for the optimal threshold. Regarding the F1-measure, LogMap achieved the best score. LogMap and Matcha provide the correspondences with real-valued confidence. Therefore, we applied thresholding during the evaluation. Agent-OM internally applied a threshold of 0.9 before submission of alignments.

LogMap and Matcha use weights to score each pair in their generated alignments. The weights in LogMap's alignment range from 1.0 to 0.5. There were multiple matches with the lowest weight (0.5). These mappings were a mix of correct mappings (majority) and false positives. The computed threshold for obtaining the highest F-measure is 0.5 (including) which corresponds to the mappings analysis. Therefore, LogMap achieves the highest F-measure while taking all the mappings into consideration, without threshold adjustments.

In case of Matcha, the weights of its results range between 1 and 0.6. Surprisingly, all mappings with the highest weights were false positives. The correct mapping with the lowest weight was weighted to 0.9306. This is also the computed threshold for obtaining the highest F-measure which is 0.61. Threshold 0.9306 (including) is very specific. Using a more general threshold 0.9 (including), the F-measure is just a slightly lower 0.6. Applying the thresholds, the number of correct mappings stays the same while the number of false positives lowers from 222 to 44 for threshold 0.9306 or 46 for threshold 0.9. Applying the threshold 0.9, precision improves from 0.149 to 0.459, F-measure from 0.255 to 0.6, while recall remains 0.867.

Looking at the results, it can be said that when the reason an alignment was discovered was the same name, all or at least most tools generated the mapping. LogMap and Matcha further generated some FPs based on similar strings. Agent-OM and Matcha generated some FPs based on synonyms. All four systems generated FPs where the same word was present in the entities names.

Last year, the CE track included only one ontology pair for matching, CEON and BiOnto. Comparing the results from last year,[36] all three common participants (LogMap, LogMapLt, Matcha) significantly improved their performance this year (Matcha after thresholding). Based on the false positives analysis, just as last year, it turns out that mere string matching could be misleading, and the meaning of entities should be better considered. This approach could be an opportunity for future performance improvements.

[36]More information is provided at the results web page: https://oaei.ontologymatching.org/2025/results/ce/index.html

4.11. Beyond Equivalence

Table 23 summarises the performance of the participating matchers across all 10 datasets, 5 from industrial classification standards (**ICS**) and 5 from STROMA/TaSeR (**ST**). Table 24 provide the results use *isAmong* evaluation, which provide more fine-grained measurement for correspondences beyond equivalence.

isAmong is a novel evaluation framework designed to assess ontology matching beyond simple equivalence. Unlike traditional approaches that focus solely on exact class-to-class matches, isAmong introduces a relation-aware perspective by transforming alignments into sets of descendant classes—called isAmong sets. This enables the computation of class-level Precision, Recall, and F1-Score based on the overlap between predicted and reference descendant sets, averaged across both source and target ontologies. By rewarding containment and partial overlap, isAmong provides a fair and fine-grained metric even when systems do not predict the exact reference relation. Tailored for classification ontologies, the framework avoids pre-defined weights and supports fine-grained evaluation for correspondences with relation beyond equivalence ($\equiv$), such as subclass ($\leq$), superclass ($\geq$), and overlap ($\simeq$).

Table 23
Traditional evaluation (Precision, Recall, F1) for the Beyond Equivalence track with per-group averages.

Dataset	Precision					Recall					F1-score				
	LogMap	LogMap Bio	LogMap KG	Matcha	MDMapper	LogMap	LogMap Bio	LogMap KG	Matcha	MDMapper	LogMap	LogMap Bio	LogMap KG	Matcha	MDMapper
eclass-gpc	32.35	30.56	30.56	–	**12.83**	**0.09**	0.09	0.09	–	0.19	0.17	0.17	0.17	–	**0.37**
eclass-unspsc	17.34	16.10	16.10	14.09	**11.56**	0.04	0.04	0.04	**0.03**	0.11	0.07	0.08	0.08	0.05	**0.21**
etim-eclass	40.32	88.13	70.02	–	**96.77**	24.69	24.82	24.82	–	**28.55**	30.62	38.74	36.65	–	**44.09**
gpc-unspsc	24.51	24.51	24.51	21.83	**13.73**	0.13	0.13	0.13	0.16	**0.29**	0.25	0.25	0.25	0.31	**0.56**
gpc-unspsc+	**8.49**	23.15	23.15	18.18	13.79	**0.16**	0.09	0.09	0.10	0.22	**0.31**	0.17	0.17	0.20	0.44
Avg ICS	24.20	36.49	32.06	18.03	**29.74**	5.02	5.03	5.03	0.07	**5.87**	6.29	7.88	7.47	0.14	**9.13**
g1-web	3.20	60.81	16.47	–	**88.24**	**53.64**	40.91	53.64	–	36.36	6.05	48.91	25.20	–	**51.50**
g2-diseases	51.67	**60.17**	57.14	2.50	57.45	69.77	**61.02**	62.15	70.34	7.63	59.38	**60.59**	59.54	4.83	13.47
g3-text	43.75	43.75	43.75	**39.85**	38.53	**7.35**	7.35	7.35	6.96	5.51	12.58	12.58	12.58	**11.84**	9.64
g5-groceries	20.51	27.59	27.59	23.53	**46.51**	5.13	5.13	5.13	5.13	**12.82**	8.21	8.65	8.65	8.42	**20.10**
g7-literature	64.71	78.57	78.57	84.62	**100.00**	13.41	13.41	13.41	13.41	**13.41**	22.22	22.92	22.92	23.16	**23.66**
Avg ST	36.77	54.18	44.70	37.62	**66.15**	**29.86**	25.56	28.34	23.96	15.15	21.69	**30.73**	25.78	12.06	23.67
Avg All	30.69	**45.33**	38.78	25.57	47.94	**17.44**	15.30	16.68	12.02	10.51	13.99	**19.31**	16.62	6.10	16.40

This year we evaluated five matchers, including LogMap, LogMap-Bio, LogMapKG, Matcha, and MDMapper [68, 69], across 10 datasets from two families: industrial classification standards (i.e., ECLASS–GPC, ECLASS–UNSPSC, ETIM–ECLASS, GPC–UNSPSC, GPC–UNSPSC+), and STROMA/-TaSeR (i.e., g1-web, g2-diseases, g3-text , g5-groceries, g7-literature). We report both traditional metrics (Precision, Recall, F1-Score), which reward only exact identicial correspondences, and isAmong metrics (Precision*, Recall*, F1-Score*), which also give credit for partially correct relations such as superclass, subclass and overlap.

Across all 10 datasets (macro level), LogMap leads on isAmong metrics (best F1*), while LogMap-Bio achieves the highest traditional F1-Score. For industrial classification standards, MDMapper performs best under both evaluation regimes, particularly on isAmong (top P*, R*, F1*). Overall performance is very low, likely due to the scarcity of true equivalences and the dominance of other relation types. This suggests that current matchers struggle to detect relations between concepts with differing granularity or classification perspectives. For TROMA/TaSeR, LogMap-Bio achieves the best traditional F1-Score, while LogMap leads in Recall and all isAmong metrics (Precision*, Recall*, F1-Score*).

A breakdown by dataset family provides further insight:

- **Industrial Classification Standards (5 datasets):** These datasets remain the most challenging.

Table 24
isAmong evaluation (Precision*, Recall*, F1*) for the Beyond Equivalence track with per-group averages.

Dataset	Precision*					Recall*					F1-score*				
	LogMap	LogMap Bio	LogMap KG	Matcha	MDMapper	LogMap	LogMap Bio	LogMap KG	Matcha	MDMapper	LogMap	LogMap Bio	LogMap KG	Matcha	MDMapper
eclass-gpc	4.09	4.19	4.19	–	**10.81**	1.67	1.65	1.65	–	**6.09**	1.83	1.82	1.82	–	**6.29**
eclass-unspsc	3.97	4.31	4.31	3.32	**11.86**	1.59	1.75	1.75	1.14	**5.04**	1.74	1.91	1.91	1.28	**5.56**
etim-eclass	37.00	37.15	37.15	–	**42.17**	34.17	34.35	34.35	–	**38.73**	34.44	34.60	34.60	–	**39.50**
gpc-unspsc	6.92	6.92	6.92	9.14	**16.35**	3.11	3.11	3.11	4.28	**8.75**	3.23	3.23	3.23	4.32	**9.07**
gpc-unspsc+	**2.91**	1.68	1.68	2.30	4.23	**2.07**	0.90	0.90	1.17	2.93	**1.96**	0.92	0.92	1.19	2.88
Avg ICS	10.98	10.85	10.85	3.69	**17.09**	8.52	8.35	8.35	1.65	**12.31**	8.64	8.49	8.49	1.70	**12.66**
g1-web	50.10	44.59	50.10	–	**45.96**	50.22	41.25	50.22	–	**39.20**	**48.76**	41.50	48.76	–	40.69
g2-diseases	**45.81**	44.36	43.59	28.04	11.03	**47.64**	44.55	44.25	36.68	7.64	**45.92**	43.85	43.31	28.64	8.07
g3-text	19.88	19.88	19.88	**19.79**	14.97	9.11	9.11	9.11	8.91	**6.49**	**11.56**	11.56	11.56	11.47	8.37
g5-groceries	16.71	15.70	15.70	16.98	**28.84**	14.00	13.60	13.60	14.02	**24.81**	14.24	13.80	13.80	14.32	**24.61**
g7-literature	**35.90**	29.43	29.43	30.94	27.07	**31.24**	24.77	24.77	25.31	21.47	**32.56**	26.01	26.01	26.87	22.92
Avg ST	**33.68**	30.79	31.74	23.94	25.57	**30.44**	26.66	28.39	21.23	19.92	**30.61**	27.34	28.69	20.32	20.93
Avg All	**22.33**	20.82	21.29	13.81	21.33	**19.48**	17.50	18.37	11.44	16.12	**19.62**	17.92	18.59	11.01	16.80

Under isAmong evaluation, average F1* stays below 13% for all systems. MDMapper performs best in this group (F1* $\approx$ 12.66%), but the overall difficulty highlights the incapability of existing tools in ontologies with structural and granularity differences from industry cases.

– **STROMA/TaSeR (5 datasets):** The performance here is consistently higher. LogMap obtains isAmong F1* $\approx$ 30.61%, and LogMap-Bio achieves traditional F1 $\approx$ 30.73%. MDMapper yields high precision (66.15%) under the traditional metric.

In summary, the first year of the Beyond Equivalence track demonstrates that matching beyond equivalence remains a challenging and open research problem. While relation-aware evaluation provides a more realistic assessment of system capabilities, substantial methodological advancements are required for high-quality alignment in practical applications such as matching product classification ontologies.

4.12. Knowledge Graph

This year we evaluated all participants with the MELT framework to include all possible submission formats i.e., SEALS, and Web format. First, all systems are evaluated on a very small matching task[37] (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 non-empty alignments.

The resulting alignments are available for download.[38] This year's best overall system is DogMa, which beats the baselines for the first time (0.90 F-measure) and also achieved the highest recall (0.89). Detailed results for each test case can be found on the OAEI results page of the track.[39]

[37] http://oaei.ontologymatching.org/2019/results/knowledgegraph/small_test.zip
[38] http://oaei.ontologymatching.org/2025/results/knowledgegraph/knowledgegraph-alignments.zip
[39] http://oaei.ontologymatching.org/2025/results/knowledgegraph/index.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)
DogMa	00:00:00	5	24.4	0.98 (0.98)	0.84 (0.84)	0.78 (0.78)
LogMap	00:56:43	5	19.4	0.93 (0.93)	0.80 (0.80)	0.71 (0.71)
LogMapLt	64:48:07	4	23.0	0.80 (1.00)	0.55 (0.69)	0.43 (0.54)
LSMatch	03:31:25	5	23.6	0.97 (0.97)	0.74 (0.74)	0.64 (0.64)
Matcha	01:43:25	4	27.8	0.77 (0.96)	0.67 (0.84)	0.60 (0.75)
TIM	00:34:13	5	29.2	1.00 (1.00)	0.95 (0.95)	0.90 (0.90)
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)
DogMa	00:00:00	5	79.4	0.99 (0.99)	0.96 (0.96)	0.94 (0.94)
LogMap	00:56:43	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
LogMapLt	64:48:07	4	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
LSMatch	03:31:25	5	85.6	0.73 (0.73)	0.71 (0.71)	0.69 (0.69)
Matcha	01:43:25	4	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
TIM	00:34:13	5	92.8	0.99 (0.99)	0.97 (0.97)	0.95 (0.95)
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)
DogMa	00:00:00	5	5848.8	0.91 (0.91)	0.90 (0.90)	0.89 (0.89)
LogMap	00:56:43	5	4012.4	0.90 (0.90)	0.78 (0.78)	0.69 (0.69)
LogMapLt	64:48:07	4	6653.8	0.73 (0.91)	0.67 (0.84)	0.62 (0.78)
LSMatch	03:31:25	5	5872.2	0.66 (0.66)	0.59 (0.59)	0.60 (0.60)
Matcha	01:43:25	4	29113.8	0.54 (0.67)	0.62 (0.77)	0.72 (0.90)
TIM	00:34:13	5	5243.4	0.91 (0.91)	0.88 (0.88)	0.85 (0.85)
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)
DogMa	00:00:00	5	5952.6	0.91 (0.91)	0.90 (0.90)	0.89 (0.89)
LogMap	00:56:43	5	4031.8	0.90 (0.90)	0.77 (0.77)	0.68 (0.68)
LogMapLt	64:48:07	4	6676.8	0.73 (0.92)	0.66 (0.83)	0.61 (0.76)
LSMatch	03:31:25	5	5981.4	0.66 (0.66)	0.60 (0.60)	0.61 (0.61)
Matcha	01:43:25	4	29141.5	0.54 (0.68)	0.61 (0.76)	0.71 (0.88)
TIM	00:34:13	5	5365.4	0.92 (0.92)	0.88 (0.88)	0.85 (0.85)

Property matches are still not created by all systems. LogMap and Matcha do not return any of those mappings. One reason might be that the properties are typed as `rdf:Property` and not distinguished into `owl:ObjectProperty` or `owl:DatatypeProperty`.

When it comes to class matches, TIM is the overall best system with an F-measure of 0.95 (much better than the provided baseline).

For further analysis of the results, we also provide an online dashboard[40] generated with MELT [70]. 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, LSMatch (03:31:25) and LogMapLt (64:48:07) were the slowest systems. Besides the baselines (which need around 12 minutes for all test cases) TIM (00:34:13) is the fastest system.

[40] http://oaei.ontologymatching.org/2025/results/knowledgegraph/knowledge_graph_dashboard.html

4.13. Pharmacogenomics

For this third year of the Pharmacogenomics track, no systems registered for the track. Nevertheless, we evaluated some of the systems submitted to OAEI 2025, namely LogMap (with its different versions: LogMap, LogMap-Bio, LogMapLt, and LogMapKG), LSMatch, LSMatch-Multilingual, Matcha, and TIM using the MELT framework. LSMatch, LSMatch-Multilingual, and TIM failed to produce alignments due to runtime errors.

Regarding LogMap, similarly to last year, its different versions did not produce alignments between n-ary tuples; however, some versions produced alignments between other entities (e.g., components of pharmacogenomic tuples such as drugs or genetic factors). These alignments were valid, sometimes trivial, but are out of the scope of the Pharmacogenomics track. We link the inability of LogMap to produce alignments between n-ary tuples to the absence of labels for such tuples, as providing labels allows all 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. These observations lead us to conclude that LogMap is not adequate for the task of matching pharmacogenomic knowledge, as it appears to rely only on labels and disregard neighbors.

Matcha failed to produce alignments on the three proposed tasks due to out-of-memory errors. However, when tested on sample tasks involving only two tuples to match, Matcha was able to produce alignments between these n-ary tuples, even in the absence of labels. We also noticed that altering the neighborhood of tuples may have varying impacts on the produced alignments. For instance, removing one of the two inverse relations linking one tuple to its components leads to the absence of output alignments, whereas removing the two edges for one tuple still allows alignments to be detected. As a result, we believe Matcha could be an interesting candidate for pharmacogenomic knowledge alignment, even if it would require to be adapted to tackle the huge number of tuples to align, and the diversity of their neighborhoods.

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 runtime 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 participants. Despite this fact, this year, new participants came up with alternative strategies (i.e., deep learning) with respect to the last campaigns.

The **Bio-ML** track attracted several new machine learning-based participants. However, the number of symbolic participants is still low. The best-performing systems are not consistent across tasks and settings, demonstrating the diversity of our datasets.

The results of the **Digital Humanities** track show that SKOS vocabularies are still not well-supported by many matching systems. However, there is an improvement to last year with two new systems capable of handling SKOS. For all systems, there is still room for improvement.

The **Archaeology multilingual** track leads to the conclusion that languages other than English are not well-supported. The newcomer Agent-OM showed promising results in this track. In future versions, it is planned to include other languages such as Japanese.

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 11 and 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, two systems were able to complete the task. A new dataset is an addition this year for the specific task of complex multi-ontology matching.

Automatic instance-matching benchmark generation algorithms have been gaining popularity, as evidenced by the fact that they are used in instance-matching tracks. 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, we have two new matching systems, TIM and DogMa, which beat the previous best F-Measures in class/property and instance performance. We hope that, in the future, more systems will focus on the track and continue to improve.

For the third year of the **Pharmacogenomics** track, we tested several systems submitted to OAEI 2025, even if they did not specifically register for the track. None of these systems were successful in producing alignments between reified n-ary tuples, which, according to our investigation, is due to the absence of labels for the n-ary tuples to align, or to scalability issues. In particular, Matcha is the most promising candidate as it is able to match n-ary tuples but fails due to their high number. These results highlight the interest of considering domain-specific problems that bring additional challenges to the field of ontology matching (here, different types of alignments between individuals, structure-based matching). Given the inability of registered systems to produce valid alignments, such challenges are currently unaddressed and require to design new methods like [58, 71] or enrich existing ones. This ultimately motivates to propose the track again in future editions of OAEI, hoping to motivate new systems targeting this real-world matching scenario. We will also enrich the track with tasks that involve fewer entities to match, or specific variations in tuple neighborhoods, allowing to assess even further systems capabilities and limits.

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. Since 2023, as a growing number of systems rely on Large Language Models, we started discussing the specific requirements and alternative ways of gathering the alignments generated by such resource-consuming systems. This year, it is highlighted that more systems rely on external calls to LLMs, which are not yet well adapted to run within our evaluation platforms. Hence, a platform capable of running resource-intensive systems is needed, potentially supported by funding or an infrastructure project, while complex alignments and other relations than equivalence remain to be addressed, reference alignments in the OAEI datasets should be revised, and an in-use session featuring real industrial cases (possibly as a special workshop session) would be valuable.

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).

Catia Pesquita was 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, 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].

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).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

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. doi:10.1007/978-3-642-22630-4_6.

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

[3] Y. Sure, O. Corcho, J. Euzenat, T. Hughes (Eds.), Proceedings of the 3rd International Workshop on Evaluation of Ontology-based Tools held at the 3rd International Semantic Web Conference ISWC 2004, Hiroshima, Japan, volume 128, 2004. URL: https://ceur-ws.org/Vol-128/.

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

[5] M. Abd Nikooie Pour, A. Algergawy, E. Blomqvist, P. Buche, J. Chen, A. Coulet, J. Cufi, H. Dong, D. Faria, L. Ferraz, P. G. Cotovio, Y. He, S. Hertling, I. Horrocks, L. Ibanescu, S. Jain, E. Jiménez-Ruiz, N. Karam, F. Kraus, P. Lambrix, H. Li, Y. Li, P. Monnin, H. Paulheim, C. Pesquita, A. Sharma, P. Shvaiko, M. Silva, G. Sousa, C. Trojahn, J. Vatascinova, B. Yaman, O. Zamazal, L. Zhou, Results of the Ontology Alignment Evaluation Initiative 2024, in: E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn, S. Hertling, H. Li, P. Shvaiko, J. Euzenat (Eds.), Proceedings of the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference

(ISWC 2024), Baltimore, USA, November 11, 2024, volume 3897 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 64–97. URL: https://ceur-ws.org/Vol-3897/oaei2024_paper0.pdf.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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, volume 2536 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 46–85. URL: https://ceur-ws.org/Vol-2536/oaei19_paper0.pdf.

[11] 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), volume 2288 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 76–116. URL: https://ceur-ws.org/Vol-2288/oaei18_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, 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, volume 2032 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 61–113. URL: http://ceur-ws.org/Vol-2032/oaei17_paper0.pdf.

[13] 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 Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, volume 1766 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016, pp. 73–129. URL: https://ceur-ws.org/Vol-1766/oaei16_paper0.pdf.

[14] 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 Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, volume 1545 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015, pp. 60–115. URL: https://ceur-ws.org/Vol-1545/oaei15_paper0.pdf.

[15] 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 Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda (IT), volume 1317 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014, pp. 61–104. URL: http://ceur-ws.org/Vol-1317/oaei14_paper0.pdf.

[16] 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 Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney (NSW, AU), volume 1111 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2013, pp. 61–100. URL: https://ceur-ws.org/Vol-1111/oaei13_paper0.pdf.

[17] 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 Workshop on Ontology Matching (OM-2012) collocated with the 11th International Semantic Web Conference (ISWC-2012), Boston (MA, US), volume 946 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2012, pp. 73–115. URL: https://ceur-ws.org/Vol-946/oaei12_paper0.pdf.

[18] 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 Workshop on Ontology Matching, Bonn (DE), volume 814 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2011, pp. 85–110. URL: https://ceur-ws.org/Vol-814/oaei11_paper0.pdf.

[19] 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 Workshop on Ontology Matching (OM-2010) collocated with the 9th International Semantic Web Conference (ISWC-2010), Shanghai (CN), volume 689 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2010, pp. 85–117. URL: https://ceur-ws.org/Vol-689/oaei10_paper0.pdf.

[20] 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 Workshop on Ontology Matching (OM-2009) Collocated with the 8th International Semantic Web Conference (ISWC-2009), Chantilly (VA, US), volume 551 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2009, pp. 73–126. URL: https://ceur-ws.org/Vol-551/oaei09_paper0.pdf.

[21] 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 International Workshop on Ontology Matching (OM-2008) Collocated with the 7th International Semantic Web Conference (ISWC-2008), Karlsruhe (DE), volume 431 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2008, pp. 73–120. URL: https://ceur-ws.org/Vol-431/oaei08_paper0.pdf.

[22] 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 of the 2nd International Workshop on Ontology Matching, Busan (KR), volume 304 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2007, pp. 96–132. URL: http://ceur-ws.org/Vol-304/paper9.pdf.

[23] 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 Workshop on Ontology Matching, Athens (GA, US), volume 225 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2006, pp. 73–95. URL: http://ceur-ws.org/Vol-225/paper7.pdf.

[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, volume 2288 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 49–60. URL: https://ceur-ws.org/Vol-2288/om2018_LTpaper5.pdf.

[25] 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. doi:10.1007/978-3-030-33220-4_17.

[26] 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. doi:10.1093/database/baac035.

[27] 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.

[28] 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. doi:10.1016/j.websem.2017.01.001.

[29] 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. doi:10.1016/j.websem.2012.04.001.

[30] L. Zhou, M. Cheatham, A. Krisnadhi, P. Hitzler, A Complex Alignment Benchmark: GeoLink Dataset, in: Proceedings of the 17th International Semantic Web Conference, Monterey (CA, USA), 2018, pp. 273–288. doi:10.1007/978-3-030-00668-6_17.

[31] L. Zhou, M. Cheatham, A. Krisnadhi, P. Hitzler, GeoLink Data Set: A Complex Alignment

Benchmark from Real-world Ontology, Data Intell. 2 (2020) 353–378. doi:10.1162/dint_a_00054.
[32] L. Zhou, C. Shimizu, P. Hitzler, A. Sheill, S. Estrecha, C. Foley, D. Tarr, R. D, The Enslaved Dataset: A Real-world Complex Ontology Alignment Benchmark using Wikibase, in: 29th ACM International Conference on Information and Knowledge Management, 2020. doi:10.1145/3340531.3412768.
[33] M. Silva, D. Faria, C. Pesquita, Reference Alignments for biomedical Complex Multi-Ontology Matching tasks (HP, MP, WBP), 2025. doi:10.21227/btb7-yd20.
[34] M. C. Silva, D. Faria, C. Pesquita, Complex multi-ontology alignment through geometric operations on language embeddings, in: 27th European Conference on Artificial Intelligence, ECAI 2024, IOS Press BV, 2024, pp. 1333–1340. doi:10.3233/FAIA240632.
[35] M. C. Silva, D. Faria, C. Pesquita, CMOMgen: Complex Multi-Ontology Alignment via Pattern-Guided In-Context Learning (2025). URL: http://arxiv.org/abs/2510.21656. doi:10.48550/arXiv.2510.21656, arXiv:2510.21656 [cs].
[36] G. Santos Sousa, R. Lima, C. Trojahn, On Evaluation Metrics for Complex Matching Based on Reference Alignments, in: European Semantic Web Conference, Springer, 2025, pp. 77–93. doi:10.1007/978-3-031-94575-5_5.
[37] 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. doi:10.1007/978-3-642-38288-8_3.
[38] 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. doi:10.1007/978-3-319-46523-4_13.
[39] 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.
[40] 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. doi:10.1007/978-3-319-18818-8_1.
[41] 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. doi:10.1007/978-3-031-19433-7_33.
[42] 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. doi:10.1101/2022.04.13.22273750.
[43] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.
[44] 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. doi:10.1007/978-3-642-25073-6_18.
[45] Y. He, J. Chen, H. Dong, I. Horrocks, C. Allocca, T. Kim, B. Sapkota, DeepOnto: A Python package for ontology engineering with deep learning, Semantic Web 15 (2024) 1991–2004. doi:10.3233/SW-243568.
[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. URL: https://ceur-ws.org/Vol-2518/paper-WODHSA11.pdf.
[47] 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), volume 3636 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3636/paper1.pdf.

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

[49] K. Cheung, J. Drennan, J. Hunter, Towards an Ontology for Data-driven Discovery of New Materials, in: Semantic Scientific Knowledge Integration AAAI/SSS Workshop, 2008, pp. 9–14. URL: https://cdn.aaai.org/Symposia/Spring/2008/SS-08-05/SS08-05-003.pdf.

[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), volume 3753 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3753/short1.pdf.

[52] H. Li, J. Vatascinova, O. Zamazal, Y. Li, P. Lambrix, E. Blomqvist, Results and Discussions from Aligning Ontologies in the Circular Economy Domain, in: Proceedings of the 3rd International Workshop on Knowledge Graphs for Sustainability (KG4S2025), volume 4002 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4002/paper7.pdf.

[53] P. Arnold, E. Rahm, Enriching ontology mappings with semantic relations, Data & Knowledge Engineering 93 (2014) 1–18. doi:10.1016/j.datak.2014.07.001.

[54] S. Hertling, H. Paulheim, Transformer Based Semantic Relation Typing for Knowledge Graph Integration, in: The Semantic Web, 20th International Conference, ESWC 2023, Springer, 2023, pp. 105–121. doi:10.1007/978-3-031-33455-9_7.

[55] S. Hertling, H. Paulheim, DBkWik: extracting and integrating knowledge from thousands of Wikis, Knowledge and Information Systems (2019). doi:10.1007/s10115-019-01415-5.

[56] S. Hertling, H. Paulheim, DBkWik: A Consolidated Knowledge Graph from Thousands of Wikis, in: Proceedings of the IEEE International Conference on Big Knowledge, 2018. doi:10.1109/ICBK.2018.00011.

[57] 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.

[58] 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. doi:10.1007/978-3-030-57855-8_4.

[59] 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. doi:10.1186/S12859-019-2693-9.

[60] 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. doi:10.1609/aaai.v36i5.20510.

[61] 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. doi:10.1007/s11280-023-01169-9.

[62] J. Dabrowski, E. V. Munson, 40 years of searching for the best computer system response time, Interacting with Computers 23 (2011) 555–564. doi:10.1016/j.intcom.2011.05.008.

[63] Z. Qiang, W. Wang, K. Taylor, Agent-OM: Leveraging LLM Agents for Ontology Matching, Proc. VLDB Endow. 18 (2024) 516–529. doi:10.14778/3712221.3712222.

[64] S. Oulefki, L. Berkani, N. Boudjenah, L. Bellatreche, A. Mokhtari, BioGITOM: Matching Biomedical Ontologies with Graph Isomorphism Transformer, The VLDB Journal 34 (2025) 65. doi:10.1007/s00778-025-00943-7.

[65] D. Faria, M. C. Silva, P. Cotovio, P. Eugénio, C. Pesquita, Matcha and Matcha-DL results for OAEI 2022, in: 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. 197–201. URL: https://ceur-ws.org/Vol-3324/oaei22_paper11.pdf.

[66] D. Faria, M. C. Silva, P. Cotovio, L. Ferraz, L. Balbi, C. Pesquita, Results for Matcha and Matcha-DL in OAEI 2023, 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, pp. 164–169. URL: https://ceur-ws.org/Vol-3591/oaei23_paper6.pdf.

[67] 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 co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 11th, 2024, volume 3897 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3897/om2024_LTpaper1.pdf.

[68] X. Liu, J. Grode, M. R. Hansen, MDMapper: A Framework for Aligning Master Data Models using Ontology Matching Techniques, in: Proceedings of the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 11th, 2024, volume 3897 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3897/om2024_LTpaper3.pdf.

[69] X. Liu, M. R. Hansen, J. Grode, MDMapper Results for OAEI 2024, in: Proceedings of the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 11th, 2024, volume 3897 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3897/oaei2024_paper1.pdf.

[70] J. Portisch, S. Hertling, H. Paulheim, Visual Analysis of Ontology Matching Results with the MELT Dashboard, in: The Semantic Web: ESWC 2020 Satellite Events, 2020, pp. 186–190. doi:10.1007/978-3-030-62327-2_32.

[71] 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. doi:10.3233/SW-210452.

Linköping, Lisboa, Haryana, Mannheim, London, Trento, Toulouse, Paris, Prague, Manhattan, Dublin, Grenoble, Karlsruhe, Copenhagen

December 2025