A Gold Standard Benchmark Dataset for Digital Humanities

Felix Kraus[1], Nicolas Blumenröhr[1], Germaine Götzelmann[1], Danah Tonne[1] and Achim Streit[1]

[1]*Karlsruhe Institute of Technology, Karlsruhe, Germany*

Abstract

We present a benchmark dataset specifically designed to evaluate matching systems using controlled vocabularies from the digital humanities (DH). This dataset includes manually compiled gold standard alignments for eight DH test cases, addressing DH-specific challenges such as multilingualism, specialized terminology, and the use of SKOS (Simple Knowledge Organization System) as a data model. The dataset, including the reference, is publicly and persistently available and incorporated into the OAEI 2024.

To obtain a high-quality dataset, we developed requirements including criteria for resource selection and present their practical implementation. By focusing on test cases that closely reflect real-world vocabularies, we facilitate advancements of matching systems, especially for subsequent mapping and integration tasks.

Evaluating the dataset using OAEI systems revealed significant weaknesses in their handling of SKOS and multilingual data, which shows the significance of our dataset. The evaluation also highlights the dataset's quality, validity, limitations, and lessons learned, offering valuable insights for future benchmark development. We believe this dataset will substantially benefit the advancement of matching systems not only within the DH but also in other fields.

Keywords

Ontology Matching, Controlled Vocabularies, Reference Dataset, Digital Humanities, OAEI

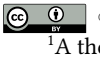
1. Introduction and Motivation

Ontologies, thesauri and controlled vocabularies (CVs)[1] play an important role in organizing and structuring knowledge. They enable researchers to use computer software to query linked data and use it for data annotation. Consequently, different ontologies and CVs developed and used by different parties in related domains lead to overlaps in content [1]. Ontology matching helps in aligning and integrating ontologies and thesauri, and is crucial for solving the heterogeneity problem. Even though the initial intention of ontology matching was to use it for full-fledged

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

felix.kraus@kit.edu (F. Kraus)

0000-0002-2102-4170 (F. Kraus); 0009-0007-0235-4995 (N. Blumenröhr); 0000-0003-3974-3728 (G. Götzelmann); 0000-0001-6296-7282 (D. Tonne); 0000-0002-5065-469X (A. Streit)



[1]A thesaurus is a special form of a controlled vocabulary that organizes terms with synonyms and hierarchical relationships. Unlike ontologies, which provide detailed frameworks for semantic reasoning, thesauri focus on standardizing terminology to enhance information retrieval. In this paper, the term *controlled vocabulary* is used as an umbrella term for thesauri and related concepts.

ontologies, its methods can also be applied to controlled vocabularies[2].

To systematically evaluate ontology matching systems, the Ontology Alignment Evaluation Initiative (OAEI) provides a platform offering a comprehensive benchmark framework as well as benchmark datasets of different domains and focus. Applying matching systems to digital humanities (DH) reveals the special challenges that are posed to the system in this domain: highly specific domain terminology leading to rather small CVs, the use of multiple (ancient) languages, low resources, and the widespread use of SKOS[3] (Simple Knowledge Organization System) as semantic web compatible data model for CVs. However, the existing benchmarks do not adequately address these challenges, which hinders progress in matching system development.

Our dataset fills this gap. It consists of eight test cases, each consisting of a CV pair and a manually compiled gold standard reference alignment. Compared to existing datasets, our resource provides several unique features: It incorporates multiple languages where the translation was done by the domain experts, and it uses SKOS as a data model. It is composed of CVs from archaeology, cultural history and the DH paired with computer science. These features lead to the fact that the dataset closely resembles real-world applications and fosters the direct use of the matching results for following tasks like merging different CVs. Finally, the reference alignments are manually compiled gold standards. This makes it possible to fully evaluate matching systems without a potential penalty for correct mappings that might be missing in the reference.

In summary, this paper presents the following contributions to the field:

- The development of DH-specific requirements for a benchmark dataset and their implementation,
- the creation of manually compiled gold standard alignments for eight DH test cases,
- the publication of the benchmark dataset with persistent URL[4] under CC-BY licence, incorporated in the OAEI 2024[5], and
- the evaluation of the quality and validity of this dataset using OAEI systems.

2. Related Work

2.1. Limitations of Existing OAEI Tracks

Although the OAEI offers a multitude of different tracks, a closer examination shows that they hold several limitations when aiming for the improvement of matching systems for the DH domain. The OAEI is largely dominated by STEM tracks, as seen in Figure 1. This is in line with the fast progress of matching systems for STEM in the past years, especially in the biomedical domain. Within the wide domain range of the OAEI, the enslaved dataset [2], as part of the complex track from OAEI 2020 to 2022 [3] is the only dataset covering humanities terms. It uses two OWL ontologies which are based on the Enslaved Project[6]. Since it is

[2]If these use a data model compatible to semantic web.
[3]http://www.w3.org/TR/skos-primer
[4]https://doi.org/10.5281/zenodo.12731589 (current version at https://github.com/FelixFrizzy/DH-benchmark)
[5]https://oaei.ontologymatching.org/2024/digitalhumanities/index.html
[6]https://enslaved.org/

monolingual and not using SKOS, it does not fulfil the described needs. There are a few other tracks that use SKOS which all come from different domains: The library track [4] (OAEI 2012 to 2014) uses two resources from economics and social sciences, which use a minimal SKOS representation. Even though there are some terms in French, the used ontologies are not semantically rich and therefore not comparable to a CV typically used in research projects, since descriptions and relations between terms are largely missing. Other tracks using SKOS CVs are the environment track [5] (OAEI 2007), the food track [6] (OAEI 2006 to 2007) and a second, older library track [5, 7, 8] (OAEI 2007 to 2009). The latter is also tackling cross-lingual matching, so does the Very Large Crosslingual Resources track [7, 8, 9] (OAEI 2008 to 2010). All of these tracks do not contain specific DH terminology and were therefore not considered for this benchmark.

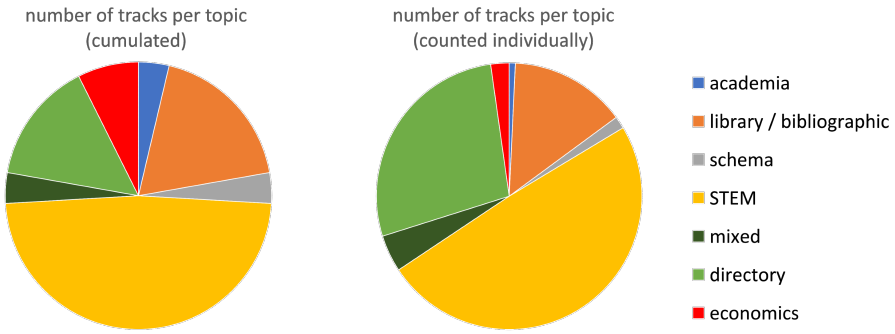


Figure 1: Overview of the topics of OAEI tracks since 2004. Left diagram: Each track counts only once, even if it was part of the OAEI in multiple years. Right diagram: Each track counted as often as it took part in any OAEI.

As shown, the few tracks that have at least some overlap with a humanities discipline were not suitable to foster DH matching system development. The same applies to the other OAEI tracks, since these also do not meet the aforementioned criteria and are too different from real-world DH CVs.

Another significant issue with existing reference alignments is the presence of incomplete or incorrect reference mapping, which leads to a bias in evaluation results. Another obstacle was the unclear provenance of some references, making it difficult to assess of their quality and reliability. Additionally, the unavailability of certain datasets severely impairs the traceability and reproducibility of evaluation results.

Lastly, the distinction between equivalence, similarity, and relatedness is often overlooked in benchmarks [10]. This distinction is crucial for creating a high-quality reference, especially when multiple languages are involved. For instance, in the biodiversity and ecology track [11], the confidence level for the alignment of the term *stellar* wind, present in both ontologies of the dataset, was only 0.85 instead of the expected value of 1 for the reference. This suggests that the reference might be created using matching systems which is not ideal.

2.2. CVs and Knowledge Graphs (KGs)

The first step in dataset creation is identifying appropriate CVs. Therefore, we present an overview of existing CVs. One way to find relevant vocabularies is to use a registry. However, to the best of our knowledge, there is currently no dedicated registry for DH vocabularies. This is why we used more generic registries such as ARDC Research Vocabularies Australia, BARTOC, CESSDA Vocabulary Service, Library of Congress Linked Data Service or the Linked Open Vocabularies, among others. The exploitation of registries provided us with well over 200 CVs within the DH.

Large KGs like Wikidata, GermaNet, WordNet or DBpedia should also be mentioned. As stated by Morvillo et al. [12], the information in large KGs is often of general nature, relevant technical terms might be missing. Nevertheless, they serve as a good starting point when building e.g. a project-specific CV, especially in combination with tools for matching entities to KGs like Mix-n-match[7].

Due to our common research activities with different DH research projects involving scholars from many fields, we noticed that the use of large knowledge graphs is often not useful because relevant technical terms are often missing in there [12]. Therefore, we decided not to exploit these data sources.

3. Approach

Our goal is to provide a benchmark dataset that closely resembles real research data. This ensures a robust foundation for further developments of ontology matching systems tailored to DH use cases, such as the matching and integration tasks of ontologies, particularly CVs. To achieve this, we use CVs from existing research projects and edit them as little as possible, preserving their original structure and content. The path to achieving this goal is described in this section. The requirements that CVs must fulfil to be integrated into our dataset are introduced gradually in the following subsections. An overview over all steps is depicted in Figure 2.

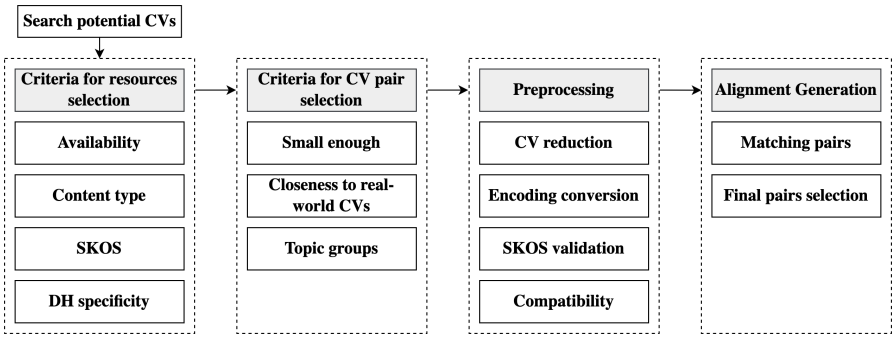


Figure 2: The steps for creating our benchmark dataset.

[7]https://mix-n-match.toolforge.org/ (source code:https://bitbucket.org/magnusmanske/mixnmatch/)

3.1. Criteria for Resources Selection

Availability We only considered resources that use a licence which allows modification and distribution. Resources without any licence information could not be used.

Content Type We did not use resources only available in plain HTML, PDF, DOCX or other data formats that are incompatible with linked data principles without further modification.

SKOS Some matching systems can only handle OWL ontologies and do not support SKOS. OWL primarily uses multiple classes, class instances and properties to specify relationships. In contrast, SKOS primarily uses a single `skos:Concept` class to define concepts ("An idea or unit of thought") and a finite set of properties. Converting SKOS to OWL might require long, hard modelling effort [13]. This was attempted in a limited scope in the OAEI 2007 library track or by Miklos Nagy[8]. This step requires expert knowledge and could detract from the dataset's real-world applicability and was avoided.

DH Specificity The CVs have to be specific to the DH domain. We ensured this by selecting CVs that include multiple languages, domain-specific terms relevant to DH research, and are of high quality, meaning that they were created by experts, usually as part of research projects [14].

Out of the initial pool of over 200 CVs that we examined, only 20 fulfilled the aforementioned requirements and were analysed.

3.2. Criteria for CV Pair Selection

After the selection process described in the previous chapter, we examined all possible $\binom{20}{2} = 190$ CV pairs using the criteria and steps described below.

CV Size As previously emphasized, creating complete alignments is of particular importance for our dataset. Since manual alignment is a laborious task as also noted by Thiéblin et al. [15], we only considered CV pairs with a manageable[9] number of potential term pairs. This means selecting CV pairs where one CV was small enough to allow going through all its terms and searching for potential matches in the other CV.

Closeness to Real-World CVs Another primary goal of our dataset is to stay close to real-world scenarios while retaining the usefulness for the OAEI. This ensures that the dataset is suited for subsequent tasks, providing practical utility for real-world applications.

Topic Groups To maximize the number of term matches between two CVs, we grouped them into different DH domains: archaeology, cultural history, the DH paired with computer science, historic books / library studies, and CVs that included a large amount of Latin terms.

Our pool of potential CV pairs further decreased to 32 after these selection steps.

3.3. Preprocessing Steps

CV Reduction In some cases, only a specific hierarchy branch of a CV contained relevant DH terms. To prevent coincidental matches with terms from other branches and to maintain a manageable size for manual alignment, we developed a Python script. This script allows the

[8] https://oaei.ontologymatching.org/2008/skos2owl.html

[9] *Manageable* cannot be quantified because the time for the manual search for term matches highly depends on the topical overlap, the hierarchy structure and the overall quality of a CV.

selection of any term of a CV, retaining only its direct parents and all ancestors within the hierarchy, while removing all other terms. This iterative process ensured that we did not lose potential topical overlaps when removing certain branches. It is published[10] under MIT licence.

Encoding Conversion Some matchers could not handle SKOS files in `turtle` encoding. Therefore, we used Skosify[11] for conversion to `RDF/XML`.

SKOS Validation We applied Skosify to check for loose terms (meaning they are not part of the hierarchy), for doublets and for SKOS model violations that were repaired. Apart from these minor corrections, the original sources were not further edited.

Compatibility A prerequisite for the CVs was that they can be parsed with both OwlApi (v5.1.19) and Apache Jena (v3.12.0), as matching systems usually use either of these two for data handling. Since Apache Jena expects at least one OWL class, (see subsection 3.1 for the differences of OWL and SKOS), we added the `skos:Concept` class manually to each SKOS file, see Figure 3 in Appendix A.

Pruning / Filtering Since we targeted smaller CVs, there was no need to apply more advanced pruning (proposed by He et al. [16]) or filtering techniques (proposed by Fallatah et al. [17]).

3.4. Alignment Generation

Term Match Definition We want to clarify our understanding of *term match* because this remains ambiguous in several other dataset descriptions, as pointed out by Hill et al. [10]:

- Two terms are considered a match if they are semantically equivalent, in other words, they have the same meaning.
- Consequently, we only consider the same part of speech as a match. This is crucial because zero derivation[12] is common in English, unlike in other languages. For example, the noun *attack* and the verb *to attack* would be incorrectly matched by a simple string matcher, whereas their German translation *Angriff* and *angreifen* are distinct in spelling.

Relations We chose not to exploit relations in the source CVs like `skos:exactMatch` due to the dependence on the creator's accuracy, whose identity often remains unclear. Nevertheless, we did not remove these relations from the sources to allow for a more thorough evaluation of matching systems, which might exploit these (potentially inaccurate) relations.

Finding Matching Pairs To identify term matches, we went through each term in the smaller CV of a CV pair and searched for equivalent terms in the other. To achieve this, we used full-text search and exploited the hierarchy. The latter means that we were looking for an equivalent term in the hierarchy branches where we would expect them topic wise. Since we use high-quality CVs with well-structured hierarchies, we are convinced that our reference alignments are gold standards and well suited for matching system evaluation.

Alignment Format We used the Alignment format[13] developed for the Alignment API. EDOAL, the extension of the Alignment format, allows representing complex alignments, with

[10] https://github.com/FelixFrizzy/rdf-tools/tree/main/hierarchy-subranches
[11] https://github.com/NatLibFi/Skosify
[12] Zero derivation means creating a new word in another part of speech from an existing word.
[13] https://moex.gitlabpages.inria.fr/alignapi/format.html

the downside of difficulties for humans to parse [2]. We prioritized readability and therefore used the Alignment format.

Terms without matches In our test cases, most terms have no match in the corresponding CV. We are interested in false positives generated by matching systems, so we did not apply additional measures to remove all terms without any match, in contrast to Fallatah et al. [17]. Removing these reduces the transferability of evaluation results to subsequent real-world tasks. Within the OAEI, a related issue is the assumption that at all times, alignments can be found, which is in real-world scenarios not always the case [18]. Therefore, it is essential to minimize false positives in the matching system results such that they accurately reflect such cases.

Final CV Pair Selection We selected pairs suitable for manual alignment, focusing on those with a considerable number of term matches. This resulted in eight pairs built from nine distinct CVs. Historic books, library studies and Latin CVs were not included in the benchmark because they were too small and had too few term matches. The CVs that we did not use are listed in Appendix A. The properties of the CVs are presented in Table 1.

Table 1
Controlled vocabularies used for the dataset.

Resource	Field[14]	Version / Date	#concepts[15]	language (ISO 639)
DEFC Thesaurus[16]	Archaeology	-	~800	de, en, la
PACTOLS thesaurus for archaeology[17]	Archaeology	- / 2021-05-18	~60,000	ar, de, en, es, fr, it, nl
Iron-Age-Danube thesaurus[18]	Archaeology	1 / 2018-11-07	~6900	de, en, hr, hu, sl
iDAI.world Thesaurus[19]	Arch. / cult. hist.	1.2 / 2022-02-10	~290	de, en, es, fr, it
PARTHENOS Vocabularies[20]	Arch. / cult. hist.	- / 2019-05-07	~4200	en
OeAI Thesaurus - Cultural Time Periods[21]	Cultural history	1.0.0 / 2022-11-23	~400	de, en
DHA Taxonomy[22]	DH/CS	- / 2018-04-03	~120	en
UNESCO[23]	DH/CS	- / 2024-06-03	~4500	ar, en, fr, es, ru
TaDiRAH[24]	DH/CS	2.0.1 / 2021-07-22	~170	de, en, es, fr, it, pt, sr

3.5. Dataset Properties

The characteristics of the source and target CVs and of the reference alignment are shown in Table 2. All CVs are in SKOS using RDF/XML as RDF syntax.

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

Table 2
Dataset Properties

Domain	Source (#terms[25])	Target (#terms)	#True Positives
Archaeology	DEFC (800)	PACTOLS (70)	11
	iDAI (2600)	PACTOLS (70)	18
	Iron-Age-Danube (290)	PACTOLS (70)	6
	PACTOLS (70)	PARTHENOS (800)	13
Cultural History	iDAI (270)	PARTHENOS (200)	53
	OeAI (400)	PARTHENOS (200)	48
DH / CS	DHA (115)	UNESCO (490)	12
	TaDiRAH (170)	UNESCO (490)	16

3.6. Specific Challenges in Dataset Construction

While building the dataset, we encountered several challenges. One unfortunate finding is that numerous resources are no longer available. A particular example is DM2E [19] which reflects the specific requirements that come from the domain of manuscripts and old prints. Although it would likely be well-suited to be included in our dataset, its unavailability prevents this. The loss of such datasets is particularly regrettable because creating ontologies and controlled vocabularies involves a significant amount of work. Consequently, future research cannot benefit from these efforts any more. Fortunately, the implementation of the FAIR (findable, accessible, interoperable, reusable) principles, especially the persistent provision of datasets, is becoming more prominent in research projects. Our dataset is available persistently and represented as a FAIR Digital Object[26] based on the concept described by Schultes et al. [20]. This ensures that future efforts can directly benefit from our work.

Another, albeit unsurprising, observation is the presence of errors in some utilized resources. We discovered duplicates, copy-and-paste errors e.g. in descriptions, and loose concepts. As described in subsection 3.3, Skosify is an excellent tool to mitigate violations of the SKOS data model. Regarding the search for suitable DH CVs, we faced the challenge of having a fairly large number of CVs but only a few with significant topical overlap. Another anticipated problem was the difficulty of involving domain experts for domain or even project specific terminology.

4. Evaluation, Preliminary Results and Discussion

To evaluate the dataset and obtain preliminary results, we used the MELT framework[27] which is also used in OAEI campaigns. We used the snapshot of the main branch of the GitHub repo from July 2024. At the time of our evaluation, the systems from OAEI 2023 were not yet available, so we used the systems from 2022[28]. Our evaluation is based on the well-established criteria macro F1-score, macro precision, macro recall and runtime. We intentionally used hardware that is close to the ones used in research projects, especially in the DH field, where cloud computing

[26] https://hdl.handle.net/21.11152/a3f19b32-4550-40bb-9f69-b8ffd4f6d0ea
[27] https://github.com/dwslab/melt
[28] https://tinyurl.com/public-oaei-systems

infrastructure or high-performance systems are often not available. Our system is equipped with an Apple M1 chip (max. 3.20 GHz), 16 GB of RAM and an SSD drive.

The results of all systems that found alignments are presented in Table 3. An overview of the runtime of these systems is provided in Table 4 in Appendix A. The following systems produced only errors and did not output any results: Matcha, ALIN, AMD, SEBMatcher and WomboCombo. Furthermore, ALIOn, ALOD2vec, CIDER-LM, LogMapLite, LSMatch, LSMatch-Multilingual and Wiktionary matcher ran without errors but did not find any alignments. We assume that most of these issues are due to incompatibility with SKOS. These might be solved during the upcoming OAEI campaign, and therefore, we did not investigate further.

Table 3
Matching system performance. The numbers are rounded to two decimal places. The highest values among the matchers are marked bold for each testcase.

Test Case	Precision					Recall					F1-score				
	AML	AT Matcher	Log Map Bio	Log Map KG	Log Map	AML	AT Matcher	Log Map Bio	Log Map KG	Log Map	AML	AT Matcher	Log Map Bio	Log Map KG	Log Map
defc-pactols	0.90	**1.00**	0.20	0.90	0.33	0.90	0.80	0.20	0.90	**1.00**	**0.90**	0.89	0.20	**0.90**	0.50
idai-pactols	**0.41**	0.31	0.40	0.40	0.35	0.41	0.24	0.71	0.71	**1.00**	0.41	0.27	0.51	0.51	**0.52**
ironage...-pactols	**0.67**	**0.67**	0.40	0.40	0.31	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**0.73**	**0.73**	0.53	0.53	0.44
pactols-parthenos	**0.83**	**0.83**	0.71	0.71	0.42	0.83	0.83	0.83	0.83	**0.92**	**0.83**	**0.83**	0.77	0.77	0.58
idai-parthenos	**1.00**	**1.00**	**1.00**	**1.00**	0.70	0.21	0.17	0.17	0.17	**0.27**	0.35	0.30	0.30	0.30	**0.39**
oeai-parthenos	0.90	0.88	**1.00**	**1.00**	0.51	0.74	0.60	0.68	0.68	**0.89**	**0.81**	0.71	**0.81**	**0.81**	0.65
dha-unesco	0.05	**0.67**	0.50	0.50	0.25	0.40	0.40	0.40	0.40	**0.90**	0.09	**0.50**	0.44	0.44	0.39
tadirah-unesco	0.50	**0.70**	0.00	0.53	0.22	0.67	0.47	0.00	0.67	**0.80**	0.57	0.56	0.00	**0.59**	0.35
Average of all tracks	0.66	**0.76**	0.53	0.68	0.39	0.62	0.54	0.47	0.64	**0.82**	0.59	0.60	0.45	**0.61**	0.48

Generally, the preliminary results show that our dataset is diverse and well-balanced, ensuring it is neither too easy nor too difficult, and certainly not trivial.

4.1. System Comparison

Precision Among the evaluated systems, ATMatcher showed the highest overall precision. However, the performance of all systems varied significantly across different test cases. On the other end of the spectrum, LogMap had the lowest precision, indicating that while it can identify a larger number of matches, a significant proportion of these are incorrect. Generally, precision is important because low precision leads to many false positives, hiding the correctly identified matches and leading to a frustrating user experience.

Recall LogMap outperformed the other systems in recall across all test cases. This suggests that LogMap is highly effective in identifying many relevant matches, although this comes at the cost of precision. High recall is beneficial in scenarios where identifying as many relevant matches as possible is crucial, even if it includes some incorrect matches. Recall is essential because low recall results in many matches being missed by the system, diminishing its value to the user and leading to substantial manual effort despite using the system. LogMap might be best suited in this regard since Humanities users value recall over precision [21].

F1-Score The F1-Score, which balances precision and recall, showed that AML, ATMatcher, and LogMapKG achieved the best scores. Similar to precision, the performance in terms of the F1-Score also varies depending on the test case. For mapping or integration tasks, a high F1-score is important to get results that benefit the user.

Coupling of Precision and Recall The observation of high precision coupled with low recall in some test cases aligns with the expectation that if fewer matches are found, those identified are more likely to be correct.

Runtime AML had a total runtime of 59s for all test cases, almost 20 times longer than ATMatcher, the fastest system with 3s runtime. Given their comparable F1-scores, ATMatcher is preferred over AML since low runtime is important for usability. The LogMap family with 15s runtime offers a good balance if high recall is needed and fast runtime is not the primary focus.

4.2. Test Case Comparison

The easiest tasks were oeai-parthenos and pactols-parthenos, which are from two different domains. At first surprising, idai-parthenos is the most challenging task, despite also using PARTHENOS as target CV and being from the same domain as OeAI. The key difference is language: iDAI is the only one that uses German as main language, with about two-thirds of terms lacking an English translation. PARTHENOS, on the other hand, uses only English. This suggests that the matching systems either do not include or only have a basic translation step. The correctly identified matches are in most cases terms that are identical in both languages.

4.3. Other Findings

As mentioned, multiple matching systems still cannot handle SKOS, which was already the case in the early library tracks more than 15 years ago [7]. Since SKOS is widely used in research, not just in the DH field, handling SKOS files is particularly important for subsequent tasks. Therefore, efforts should be made to ensure future systems and, if possible, existing systems are SKOS-compatible.

In addition, we identified some language-related false positives that are particularly remarkable. There was a false positive match between the English term *re-use* and the German term *Reuse* (a fish trap). This makes clear that simple string matching, often applied as an initial step in some matching systems, is ineffective and misleading across different languages.

5. Limitations and Lessons Learned

5.1. Alignment Creation

The primary challenge in creating manual alignments is determining whether two terms are semantically equivalent. We still have numerous CVs in reserve that require domain experts for manual alignment. To simplify this process, especially for individuals without prior experience with CVs, SKOS, or linked data in general, a software to support this process would be of extraordinary help. Ideas for such a software were proposed by Meilicke et al. [22] and extended by some aspects from Thiéblin et al. [15]. Mix-n-match might be suitable for this task.

5.2. Implementation of DH requirements

While issues like the unavailability of data sources and of domain experts remain unsolved for now, our approach demonstrates that a high-quality benchmark dataset can still be developed. Although the steps were specifically created for the DH dataset, most parts of it can be used as a generalized approach for developing a dataset for other domains. This is especially true for data sources that use SKOS.

5.3. Dataset limitations

Small CVs Even though the small size of the used CVs resembles real-world applications, it limits the evaluation of matchers that depend on large numbers of classes or instances. In turn, this dependency itself can be understood as a limitation of such matchers.

Evaluation Bias There is potential bias due to the focus on only three domains, which may not represent the diversity of real-world scenarios. This is particularly true since we cannot cover the entire field of Digital Humanities, which is not our intention anyway. Additionally, it is impossible to factor in all possible (future) application scenarios of such systems, which is why the bias cannot be fully mitigated.

Languages CVs with ancient languages such as Latin or Ancient Greek could not be used because of their scarcity and the lack of expertise for manual alignment.

Sparse Number of True Positives Some test cases have only a few matches. While this reflects real-world conditions, it poses challenges for comprehensive evaluation. It is important to note that the relevance lies in the percentage of identified alignments from all terms in a vocabulary, not the percentage of matches from all possible term pairs. If, for example, 10% of the terms in a vocabulary can be aligned with another vocabulary, it highlights valuable reuse potential for researchers, promoting reuse over reinvention.

5.4. Evaluation limitations

- Not all systems could be tested, limiting the comprehensiveness of the evaluation.
- The confidence levels of matchers' alignments were not examined in depth.
- The matching system alignments were not directly applied to a subsequent task like merging, limiting the practical assessment of the system.
- Only 1:1 matching was considered. While crucial for ensuring correctness as a preliminary step for CV merging, this does not cover complex matching scenarios [16].

6. Conclusion and Outlook

We introduced a benchmark dataset specifically designed for the DH domain, with the primary goal of advancing matching system development. The dataset addresses challenges characteristic for the DH, such as multilingualism, smaller CVs, specialist terms and the extensive use of SKOS as data model. By focusing on test cases close to real-world matching tasks, this benchmark provides a realistic and robust base for evaluation.

The resource construction for this benchmark dataset used several DH CVs as base. Criteria and their practical implementation were developed to ensure a high-quality outcome. The final manual alignment resulted in eight gold standard test cases, each consisting of a source and target CV and the reference alignment. The test cases cover the domains of archaeology, cultural history and DH / computer science. The benchmark dataset is intended as DH track within the OAEI 2024, promoting its use within the research community and providing a platform for testing and refining matching systems.

Evaluation improvements could focus on subsumption mappings as described by He et al. [16], or compare the effect of different confidence thresholds, as proposed by Zhou et al. [2]. For further evaluation focusing on multilingualism, we created a second track using the idai-pactols test case as base. In this track, we removed all languages but one from the CVs. To achieve this, we developed an MIT-licensed Python script[29]. All possible different language combinations (English, French, German, and Italian) are compiled into 10 different test cases. This track is also part of the OAEI[30] and published[31] persistently under a CC-BY licence.

To further improve matching systems, developments should focus on supporting SKOS and multilingual data, where the benchmark dataset can fully leverage its strengths. Reviving older tracks that use SKOS could once again shift the focus to the matching of SKOS vocabularies.

Additionally, a subsequent merging task of two CVs included in the evaluation would improve the transferability of the results. These tasks could also involve the prospective use of matching systems within vocabulary editors such as VocBench [23] or EVOKS [24], the latter developed by the authors.

Concerning ML-based systems, a specifically designed DH dataset for this case could be a significant benefit in evaluation, as shown for the biomedical domain by He et al. [16]. Further developments of DH test cases could also involve datasets for ancient languages such as Latin or Ancient Greek to also enable matching systems for dealing with such languages.

With this dataset, we believe that we foster significant advancements in matching systems not only limited to the DH domain, but also in all other domains that use controlled vocabularies within their research.

Acknowledgments

This research was funded by the German Research Foundation (DFG)—CRC 980 Episteme in Motion, Project-ID 191249397, the research program “Engineering Digital Futures” of the Helmholtz Association of German Research Centers, and the Helmholtz Metadata Collaboration Platform (HMC).

Disclosure: Since the author’s mother tongue is not English, LanguageTool was used to improve punctuation, grammar, and spelling. In no way was any content created by this tool.

[29] https://github.com/FelixFrizzy/rdf-tools/tree/main/remove-language
[30] https://oaei.ontologymatching.org/2024/archaeology/index.html
[31] https://doi.org/10.5281/zenodo.12731599

References

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

[2] L. Zhou, C. Shimizu, P. Hitzler, A. M. Sheill, S. G. Estrecha, C. Foley, D. Tarr, D. Rehberger, The Enslaved Dataset: A Real-world Complex Ontology Alignment Benchmark using Wikibase, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 3197–3204. doi:10.1145/3340531.3412768.

[3] É. Thiéblin, M. Cheatham, C. T. dos Santos, O. Zamazal, L. Zhou, The First Version of the OAEI Complex Alignment Benchmark, in: M. van Erp, M. Atre, V. López, K. Srinivas, C. Fortuna (Eds.), Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks, volume 2180 of *CEUR Workshop Proceedings*, CEUR-WS.org, Monterey, USA, 2018.

[4] D. Ritze, K. Eckert, Thesaurus mapping: A challenge for ontology alignment?, in: P. Shvaiko, J. Euzenat, A. Kementsietsidis, M. Mao, N. F. Noy, H. Stuckenschmidt (Eds.), Proceedings of the 7th International Workshop on Ontology Matching, volume 946 of *CEUR Workshop Proceedings*, CEUR-WS.org, Boston, MA, USA, 2012, pp. 248–249.

[5] J. Euzenat, A. Isaac, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, O. Sváb, V. Svátek, W. R. van Hage, M. Yatskevich, Results of the Ontology Alignment Evaluation Initiative 2007, in: P. Shvaiko, J. Euzenat, F. Giunchiglia, B. He (Eds.), Proceedings of the 2nd International Workshop on Ontology Matching (OM-2007), volume 304 of *CEUR Workshop Proceedings*, CEUR-WS.org, Busan, Korea, 2007, pp. 96–132.

[6] W. R. van Hage, M. Sini, L. Finch, H. Kolb, G. Schreiber, The OAEI food task: An analysis of a thesaurus alignment task, Applied Ontology 5 (2010) 1–28. doi:10.3233/AO-2010-0072.

[7] C. Caracciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilicke, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, Results of the Ontology Alignment Evaluation Initiative 2008, in: P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt (Eds.), Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008), volume 431 of *CEUR Workshop Proceedings*, CEUR-WS.org, Karlsruhe, Germany, 2008.

[8] J. Euzenat, A. Ferrara, L. Hollink, A. Isaac, C. A. 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. T. dos Santos, G. A. Vouros, S. Wang, Results of the Ontology Alignment Evaluation Initiative 2009, in: P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, N. F. Noy, A. Rosenthal (Eds.), Proceedings of the 4th International Workshop on Ontology Matching (OM-2009), volume 551 of *CEUR Workshop Proceedings*, CEUR-WS.org, Chantilly, USA, 2009, pp. 73–126.

[9] J. Euzenat, A. Ferrara, C. Meilicke, J. Pane, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, C. T. dos Santos, Results of the Ontology Alignment Evaluation Initiative 2010, in: P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, M. Mao, I. F. Cruz (Eds.), Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), volume 689 of *CEUR Workshop Proceedings*, CEUR-WS.org, Shanghai, China,

2010.
[10] F. Hill, R. Reichart, A. Korhonen, SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation, Computational Linguistics 41 (2015) 665–695. doi:10.1162/COLI_a_00237.
[11] N. Karam, A. Khiat, A. Algergawy, M. Sattler, C. Weiland, M. Schmidt, Matching biodiversity and ecology ontologies: Challenges and evaluation results, The Knowledge Engineering Review 35 (2020) e9. doi:10.1017/S0269888920000132.
[12] A. Morvillo, M. Mecella, Integrating multiple knowledge graphs in Digital Humanities, in: ST4DM 2024: Semantic Technologies for Data Management, Twente, Italy, 2024.
[13] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, E. Summers, Key Choices in the Design of Simple Knowledge Organization System (SKOS), Journal of Web Semantics 20 (2013) 35–49. doi:10.1016/j.websem.2013.05.001. arXiv:1302.1224.
[14] B. Haslhofer, A. Isaac, R. Simon, Knowledge Graphs in the Libraries and Digital Humanities Domain, in: S. Sakr, A. Zomaya (Eds.), Encyclopedia of Big Data Technologies, Springer International Publishing, Cham, 2018, pp. 1–8. doi:10.1007/978-3-319-63962-8_291-1.
[15] E. Thiéblin, M. Cheatham, C. Trojahn, O. Zamazal, A consensual dataset for complex ontology matching evaluation, The Knowledge Engineering Review 35 (2020) e34. doi:10.1017/S0269888920000247.
[16] 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, M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d'Amato (Eds.), The Semantic Web, volume 13489 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2022, pp. 575–591. doi:10.1007/978-3-031-19433-7_33.
[17] O. Fallatah, Z. Zhang, F. Hopfgartner, A Gold Standard Dataset for Large Knowledge Graphs Matching, in: Proceedings of the 15th International Workshop on Ontology Matching, volume 2788, CEUR Workshop Proceedings, 2020, pp. 24–35.
[18] S. Hertling, H. Paulheim, The Knowledge Graph Track at OAEI, in: A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, M. Cochez (Eds.), The Semantic Web, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 343–359. doi:10.1007/978-3-030-49461-2_20.
[19] K. Baierer, E. Dröge, K. Eckert, D. Goldfarb, J. Iwanowa, C. Morbidoni, D. Ritze, DM2E: A Linked Data source of Digitised Manuscripts for the Digital Humanities, Semantic Web 8 (2017) 733–745. doi:10.3233/SW-160234.
[20] E. Schultes, P. Wittenburg, FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure, in: Y. Manolopoulos, S. Stupnikov (Eds.), Data Analytics and Management in Data Intensive Domains, Communications in Computer and Information Science, Springer International Publishing, Cham, 2019, pp. 3–16. doi:10.1007/978-3-030-23584-0_1.
[21] C. Warwick, M. Terras, J. Nyhan (Eds.), Digital Humanities in Practice, UCL Centre for Digital Humanities, London, 2012.
[22] C. Meilicke, H. Stuckenschmidt, O. Šváb-Zamazal, A Reasoning-Based Support Tool for Ontology Mapping Evaluation, in: L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, E. Simperl (Eds.), The Semantic

Web: Research and Applications, volume 5554, Springer Berlin Heidelberg, Heraklion, Greece, 2009, pp. 878–882. doi:10.1007/978-3-642-02121-3_74.

[23] A. Stellato, A. Turbati, M. Fiorelli, T. Lorenzetti, E. Costetchi, C. Laaboudi, W. V. Gemert, J. Keizer, Towards VocBench 3: Pushing Collaborative Development of Thesauri and Ontologies Further Beyond, in: 17th European Networked Knowledge Organization Systems (NKOS) Workshop, volume 1937, CEUR Workshop Proceedings, Thessaloniki, Greece, 2017, pp. 39–52.

[24] F. Ernst, L. Frank, G. Götzelmann, EVOKS - Benutzerfreundliche Erstellung kontrollierter Vokabulare für die Geisteswissenschaften, in: FORGE 2023 - Forschungsdaten in Den Geisteswissenschaften: Anything Goes?! Forschungsdaten in Den Geisteswissenschaften - Kritisch Betrachtet. Konferenzabstracts, Tübingen, Germany, 2023. doi:10.5281/zenodo.8386468.

A. Appendix

The following CVs were examined, but not used:

- FISH Archaeological Objects Thesaurus
  https://skosmos.bartoc.org/49/
- European Language Social Science Thesaurus (ELSST)
  https://thesauri.cessda.eu/elsst-4/
- Art and Archaeology
  https://skosmos.loterre.fr/27X/
- Litterature
  https://skosmos.loterre.fr/P21/
- AGROVOC Multilingual Thesaurus
  https://agrovoc.fao.org/browse/agrovoc/
- Humanities and Social Science Electronic Thesaurus (HASSET)
  https://hasset.ukdataservice.ac.uk/hasset/
- DYAS Humanities Thesaurus
  https://vocabs.dariah.eu/dyas/
- Language of Bindings Thesaurus (LoB)
  https://isl.ics.forth.gr/bbt-federated-thesaurus/Language_of_Bindings/en/
- Humord
  http://data.ub.uio.no/skosmos/humord/en/
- CodiKOS
  https://web.archive.org/web/20170622142205/https://github.com/JochenGraf/CodiLab/blob/master/CodiKOS.html
- Thesauri & Ontology (THOT) for documenting Ancient Egyptian Resources
  http://thot.philo.ulg.ac.be/thesauri.html

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
>
<rdf:Description rdf:about="http://www.w3.org/2004/02/skos/core#Concept">
     <rdfs:label xml:lang="en">Concept</rdfs:label>
     <rdfs:isDefinedBy rdf:resource="http://www.w3.org/2004/02/skos/core"/>
     <skos:definition xml:lang="en">An idea or notion; a unit of thought.</skos:definition>
     <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
</rdf:Description>
```

Figure 3: The prefixes and the `skos:Concept` class were added to each SKOS file if required.

Table 4
Runtime of the matchers for each test case.

Test Case	AML	ATMatcher	LogMapBio	LogMapKG	LogMap
test case	07s	02s	04s	04s	05s
defc-pactols	08s	01s	03s	02s	02s
idai-pactols	07s	< 01s	01s	01s	01s
ironage...-pactols	07s	< 01s	01s	01s	01s
pactols-parthenos	07s	< 01s	01s	01s	< 01s
idai-parthenos	07s	< 01s	01s	01s	01s
oeai-parthenos	07s	< 01s	01s	01s	01s
dha-unesco	09s	< 01s	03s	03s	03s
tadirah-unesco	07s	< 01s	02s	02s	02s
avg over all tracks	**59s**	**03s**	**15s**	**14s**	**14s**