Matching Multiple Ontologies to Build a Knowledge Graph for Personalized Medicine

Marta Contreiras Silva, Daniel Faria, and Catia Pesquita

LASIGE, Dep. de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal

Abstract. A rich biomedical knowledge graph can support the multi-domain data integration necessary for the application of Artificial Intelligence models in personalised medicine. Constructing such a knowledge graph from already available biomedical ontologies relies on ontology matching, however, current ontology matching systems are geared towards the alignment of pairs of ontologies of the same domain one at a time. This approach, when applied to a multi-domain problem such as personalised medicine in an all vs. all fashion, poses scalability issues while also ignoring the particularities of the multi-domain aspect.
In this work we evaluate a state-of-the-art ontology matching system, AgreementMakerLight, in the task of building a network of 28 integrated ontologies to construct a knowledge graph for Explainable AI in personalised oncology, highlighting its shortcomings. To address them, we have developed a novel holistic ontology alignment strategy building on AgreementMakerLight that clusters ontologies based on their semantic overlap measured by fast matching techniques with a high degree of confidence, and then applies more sophisticated matching techniques within each cluster. We implemented two within cluster alignment strategies, one based on pairwise alignment and another on incremental alignment.
The within-cluster incremental alignment reduced alignment time by 80% when compared with within-cluster pairwise alignment, achieving 88% coverage of its mappings. Compared to an all vs. all pairwise approach, holistic approaches reduce total running time by up to 60%.
Type of submission: In Use Technology Paper

Keywords: Ontology matching, Holistic Ontology Matching, Biomedical Ontologies, Knowledge Graphs

1 Introduction

Data-centric approaches like personalized medicine have taken the forefront in biomedical research, driven by the increasing availability of biomedical data. Artificial Intelligence (AI) is positioned as a promising solution to handle these large heterogeneous datasets composed of various types of data (e.g. genomic, clinical and image data). However, the evolution of AI has favored black-box approaches that, while effective, do not foster user trust or understanding—aspects which are critical in personalized medicine, as it often involves life-or-death decisions.

To address this limitation of black-box AI approaches, there have been renewed efforts towards developing explanatory mechanisms for AI [4]. Frequent among the approaches proposed for explainable AI (XAI) is the use of Knowledge Graphs (KG), which comprehensively encode the knowledge in a domain, and can be leveraged to support user-friendly explanations when used in concert with AI methods [16, 3]. The challenge is that, the more complex the domain, the more complex and comprehensive will be the KG needed to support XAI in that domain, and few domains approach the complexity of personalized medicine.

The area of personalized medicine deals with knowledge stemming from many specific subdomains that interact in various ways, ranging from molecules (e.g. chemical compounds, genes, and proteins) to clinical and demographic factors[10]. Accordingly, ontological representations of these domains have been the subject of intense investigation. In BioPortal[19], an online repository of biomedical ontologies, there are currently more than 800 ontologies (totalling almost 9 million classes). Some of these ontologies are designed and developed as community efforts that function as community-approved representations of reality, while others are developed by single research teams and serve a more specific and localized purpose. Thus, rather than build a KG from the ground up, we posit that one can harness the wealth of publicly available knowledge through ontology matching to build the ontological layer of a KG for personalized medicine[23].

The KATY project[1] aims to develop an AI-empowered personalized medicine system to assist medical professionals and researchers in diagnosing patients more accurately, making predictions about their future health, and recommending better treatments. KATY will tackle the challenge of translating AI-based suggestions into practical decision-making processes and treatment strategies that clinicians can understand and trust by combining high performing black-box machine learning approaches with a comprehensive knowledge graph. The KG will serve as input to AI methods (e.g. directly, through embeddings) as well as encode the AI outcomes themselves to create a shared semantic space for data, scientific context and predictions capable of supporting explanation methods[31].

In a preliminary step, a careful selection of ontologies that span the domain of interest was conducted, as the goal is to reconcile the ontologies into a single cohesive knowledge model, through ontology matching techniques, to form the backbone of the knowledge graph. This resulted in a catalogue of relevant ontologies and controlled vocabularies which comprises 78 ontologies, of which 16 are referenced directly from the public data resources, and the remaining 62 were selected from our survey of BioPortal. Of these 78 ontologies, 28 were considered core to the KATY project, and the remaining 50 were considered potentially relevant, to be used only if the coverage of the 28 core ontologies is found insufficient when integrating datasets into the KATY knowledge graph.

This paper describes the process of matching the 28 core ontologies to build an integrated semantic backbone for the knowledge graph, focusing on finding simple equivalence mappings between pairs of entities belonging to the set of

[1] http://katy-project.eu/

ontologies. We further detail the requirements for ontology matching in this application, discuss the challenges found when applying a state of the art ontology matching system, and present a novel approach for holistic ontology matching that builds on an existing system, AML[9], addressing the requirements and challenges in biomedical ontology holistic matching. We performed a series of experiments to demonstrate the impact of the holistic approach and measure improvements over the baseline state of the art system.

2 Challenges in holistic biomedical ontology matching

Ontology matching (or alignment) is the process of establishing mappings (or correspondences) relating the entities (classes, properties or individuals) of two ontologies with overlapping domains. A mapping is usually represented as a tuple $< e_1, e_2, r, c >$ where e_1 and e_2 are entities of the two ontologies, r is the semantic relation between them (e.g. $\equiv, \geq, \leq, \perp$) and optionally c is a confidence score indicating how certain about the mapping is the person or algorithm who produced it [7]. A collection of mappings between two ontologies is called an alignment, and is typically stored in a file external to the ontologies, in the Alignment RDF format[2] that is the *de facto* standard in the field.

Matching biomedical ontologies is a challenging task on its own [8], both in terms of computational resources (as they are typically quite large) and in terms of the richness and complexity of the information available to match them, including a substantial lexical component where homonyms and synonyms abound [25], the presence of cross-references that establish correspondences but with no formal semantics, and the presence of logical definitions which correspond to complex ontology mappings [15, 20].

Holistic ontology matching is an extension of the pairwise ontology matching process for a set $\Omega = \{O_1, ..., O_N\}$ of ontologies with N $\geq$ 2, where a final alignment A is produced between all of them [18]. The basic approach to do this consists of uniting the alignments between all pairwise combinations of the ontologies to align, which is evidently a sub-optimal strategy computationally as in implies performing a quadratic number of ontology alignment steps.

This holistic matching challenge has been recognized by the ontology matching, schema matching and linked data communities [22, 27], with strategies to address it being usually based on exploring two different concepts: partitioning the search space in groups within which pairwise alignment is employed [11] or applying incremental matching according to a predefined order [30, 13]. Gruetze *et al.* [11] proposed grouping of linked data concepts by topic using Wikipedia and then running an alignments only between concepts in the same group. Saleem *et al.* [30] developed a method to incrementally create an integrated schema encompassing all input schema trees, by first clustering the nodes based on linguistic label similarity and then applying a tree mining technique. Hertling *et al.* [13] analyzed the impact of the ordering of ontologies in linear executions of alignments to produce an alignment of multiple ontologies and demonstrated that

[2] https://moex.gitlabpages.inria.fr/alignapi/format.html

near-optimal results can be achieved with linear efforts. Orthogonally to these works, Megdiche *et al.* [18] developed an approach based on linear programming that is able to find a stable alignment between multiple ontologies independently of the order of alignment tasks.

Building an integrated KG containing multiple ontologies requires, not only holistic matching to produce an external alignment between them, but actually merging the ontologies. Osman *et al.* [21] categorized ontology merging works according to whether they are applied after all pairwise alignments are found [1, 5] or integrated into an incremental matching approach starting from a seed ontology [2, 32].

Thus, aligning and integrating the 28 selected KATY ontologies to form the backbone of a KG for precision oncology requires tackling challenges at these three levels: biomedical ontology matching, holistic ontology matching, and holistic ontology integration. Moreover, it also requires addressing requirements in terms of quality, coverage and scalability.

Ensuring **high quality mappings** between the ontologies is a strong requirement for a system that must work with a minimal human involvement due to the size of the task, but has a high-stakes target application in healthcare. Alignments need to achieve both high precision and high recall, since both types of errors can compromise XAI approaches, either by proposing wrong explanations or not finding suitable ones.

Achieving a sufficient **coverage of all domains** in personalized oncology is mandatory to make sure that all data required to train the AI models is well-described according to domain ontologies in a way that supports building explanations. The integration of molecular and clinical data is the key to personalized medicine, which seeks to understand the play between genotype, phenotype and environment and how it bears on the effectiveness of treatments or the prognosis of diseases. This aspect requires that not only the KG covers multiple domains but that it also includes sufficient granularity.

Finally, **scalability** must also be considered. Matching 28 ontologies means that there are nearly 1.2 million classes plus their associated properties and individuals that need to be processed. Moreover, since ontologies evolve and new relevant data may be added to the KG, the time required to build the network of ontologies should not be a limiting factor in updating the system.

To build a high quality network of biomedical ontologies we need to strike a balance between quality, coverage and scalability. Filtering out lower quality mappings may result in lower coverage, higher coverage requires the ability to align more ontologies, but more sophisticated ontology matching algorithms that are able to produce higher quality and higher coverage alignments are harder to scale.

3 Enhancing AML for holistic ontology matching

3.1 AgreementMakerLight

AgreementMakerLight (AML) is an automated ontology matching system predicated on the design principles of scalability and extensibility [9]. It has been one of the best performing systems in the yearly Ontology Alignment Evaluation Initiative (OAEI) for the past eight years, excelling particularly in tracks involving biomedical ontologies [26]. This is thanks to features such as the weighting system it uses to differentiate labels and synonyms enabling fine-grained lexical matching, or its use of cross-references and logical definitions (which are singular to the biomedical domain) [8]. Given the stellar performance of AML's matching algorithms in biomedical ontology matching, using it as the baseline matching system is an added guarantee of the quality of the produced mappings.

Of note, AML's lexical matching, cross-reference matching and logical-definition matching algorithms are all implemented using a hash-search strategy that means they run in linear time [8], and therefore can be used for profiling the suite of ontologies to match with regard to their overlap, in a holistic matching scenario.

However, like all matching systems participating in the OAEI, AML is only prepared to perform pairwise matching of ontologies, and produces ontology alignments that are external to the ontologies, in the Alignment RDF format. It doesn't include the functionality of integrating ontologies through their alignment, which would be required to build a KG automatically through ontology matching. Indeed, building a KG automatically through the alignment of multiple ontologies is beyond the state of the art in ontology matching evidenced by the OAEI.

3.2 Extensions to AML

Setting aside the possibility of matching multiple ontologies simultaneously, and contemplating only the scenarios of pairwise matching or incremental matching, the only core functionality missing from AML for holistic ontology matching is the ability to merge two ontologies through their alignment, with a *simple merge*, as defined by [21].

We extended AML by implementing the functionality of converting an RDF alignment into an OWL ontology that imports the aligned ontologies and adds the axioms corresponding to the mappings: *equivalentClass* or *subClass* for equivalence or subsumption mappings between classes; *equivalentProperty* or *subProperty* for equivalence or subsumption mappings between properties; and *sameIndividual* axioms for equivalence mappings between individuals. To enable the pairwise strategy, we also implemented the functionality of merging two or more ontologies (or OWL alignments) into a single ontology, which will be necessary to combine multiple pairwise alignments into a single KG. Furthermore, for both the pairwise and the incremental strategies, we implemented the functionality of merging an ontology with all its imports, as it would be unwieldy to have a knowledge graph with OWL import statements for several local OWL ontology and/or alignment files.

3.3 Implementing holistic matching strategies using AML

Using AML and the extensions detailed above, we implemented two distinct holistic matching strategies: pairwise and incremental, which can be preceded by a clustering step and applied within-cluster, or applied globally to the full suite of ontologies to match. Since the global algorithms are the same as the within-clustering algorithms in the particular case where the number of clusters is 1, we present only the more general within-cluster algorithms. The use of clustering is motivated by the fact that there are multiple near-orthogonal sub-domains in the biomedical domain, and we can isolate groups of ontologies from each sub-domain, for which performing sophisticated ontology matching against ontologies of other sub-domains would likely produce more erroneous than correct mappings.

To enable clustering, we perform an initial anchoring step for all pairwise combinations of ontologies using linear-time matching algorithms, whereby we calculate the fraction of classes of the smallest ontology of each pair that have the same URI, direct cross-references, shared cross-references, overlapping logical definitions, or equivalent labels or synonyms to classes in the largest ontology of the pair. This anchoring is substantially quicker than performing a full pairwise matching strategy, and has the objective of determining the overlap between all ontologies with a high degree of confidence. From the anchoring results, we build an affinity matrix indicating the semantic overlap between each pair of ontologies, which we use as input for spectral clustering, to define groups of ontologies with a higher level of overlap and therefore likely within the same sub-domain.

In the within-Cluster Pairwise Alignment (CPA) strategy, each pairwise combination of the ontologies in each cluster is matched then merged, then all the merged ontology pairs of a cluster are combined and merged into a KG[3], and finally the KGs of each cluster are merged into a final single KG using the anchoring algorithm, as detailed in Algorithm 1.

In the within-Cluster Incremental Alignment (CIA) strategy, the pair of ontologies within each cluster that has the greatest overlap is matched and merged, then the resulting merged ontology is matched against the next ontology in the cluster, and so on until all ontologies in the cluster have been matched. Then, anchoring is performed also incrementally between the KG produced from each cluster, to produce a final single KG. The algorithm is detailed in Algorithm 2.

In both strategies, we used AML's automatic matching with default configurations, but with no ontologies used as background knowledge, and with the alignment repair step switched off. Using ontologies as background knowledge would be nonsensical in this setting, as any ontology that could be used effectively as background knowledge source should be, in principle, included in the suite of ontologies to match (as the goal is to build a comprehensive knowledge graph) and therefore will be merged with the other ontologies which is effectively

[3] we use KG to denote the integrated network of ontologies which constitute the semantic backbone of the full fledged KG.

Algorithm 1 Within-cluster pairwise alignment (CPA)

```
input: C->O (map of clusters to ontologies)
init: CM->OM = new map of cluster to ontologies
init: KG = new list of ontologies
init: OK = new list of ontologies
for Ci in C:
    Oi = C.get(Ci)
    for j = 0 to Oi.length-1:
        for k in j+1 to Oi.length:
            A = AML.match(Oi[j],Oi[k])
            O = merge(convert(A))
            OMi.add(O)
    CM->OM.put(Ci,OMi)
    KG[i] = OMi[0]
    for j = 1 to OMi.length:
        merge(KG[i],OMi[j])
for i = 0 to KG.length-1:
    for j = i+1 to KG.length:
        A = AML.anchor(KG[0],KG[0])
        O = merge(convert(A))
        OK.add(O)
init: KGF = OK[0]
for i = 1 to OK.length:
    merge(KGF,OK[i])
output: KGF
```

Algorithm 2 Within-cluster incremental alignment (CIA)

```
input: C->O (map of clusters to sorted ontologies)
init: KG = new list of ontologies
for Ci in C:
    Oi = C.get(Ci)
    KG[i] = Oi[0]
    for j = 1 to Oi.length:
        A = AML.match(KG[i],Oi[j])
        KG[i] = merge(convert(A))
init: KGF = KG[0]
for i = 1 to KG.length:
    A = AML.anchor(KGF,KG[i])
    KGF = merge(convert(A))
output: KGF
```

equivalent to having it as a source of background knowledge. As for the choice not to perform repair, it is predicated on our desire for completeness of the alignment over coherence [24]. Furthermore, alignment repair algorithms take arbitrary choices when faced with conflicting mappings to remove, so while it is

critical to ensure the final KG is coherent, this should involve human revision to ensure the mappings removed or edited are indeed inaccurate.

4 Integrating biomedical ontologies in a personalized oncology KG

4.1 Ontologies

The goal of our study is the integration of the 28 ontologies selected to cover the personalized oncology domain into a single KG. These ontologies are listed in Table 1, together with the biomedical sub-domains they cover, which total 19, from molecular biology to drug-side effects. Taken together, these ontologies contain 1,191,785 classes, 2,634 properties and 397,535 individuals.

4.2 Alignment strategies

To integrate the 28 KATY ontologies, we compared the two holistic matching strategies, CPA and CIA. Additionally, as a reference point, we also tested the global pairwise alignment (GPA) strategy, which corresponds to a naive use of the state of the art in ontology matching (and algorithmically, as detailed in Section 3, is the same as CPA when the number of clusters is 1).

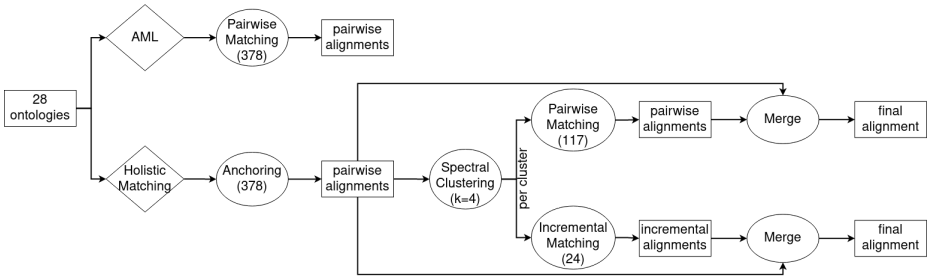

Fig. 1: Overview of the alignment strategies

4.3 Results

The global pairwise alignment (GPA) of the 28 ontologies translated into 378 alignment runs resulting in 378 pairwise alignments with a total of more than half a million mappings. The duration of the loading and matching processes[4] and total number of mappings found are presented in Table 3.

The two clustering-based approaches, CPA and CIA, require ontologies to be clustered, which involves an initial step of anchoring followed by spectral

[4] Experiments were run in a machine with 100Gb of available RAM

Table 1: Ontologies used and their domains

Acronym	Ontology	Domains	Classes
ACGT-MO	Cancer Research and Management ACGT Master Ontology	clinical feature, sample status	1769
ATC	Anatomical Therapeutic Chemical Classification	drug	6567
CCTOO	Cancer Care: Treatment Outcome Ontology	response to treatment, drug screening	1133
ChEBI	Chemical Entities of Biological Interest Ontology	metabolic, drug	171058
CL	Cell Ontology	cellular	10984
CLO	Cell Line Ontology	cell line	44873
CMO	Clinical Measurement Ontology	clinical feature, sample status	3054
DCM	DICOM Controlled Terminology	histological images	4561
DOID	Human Disease Ontology	clinical feature	17642
DTO	Drug Target Ontology	drug target interaction	10075
EFO	Experimental Factor Ontology	experimental	28816
FMA	Foundational Model of Anatomy	anatomical data	78977
GENO	Genotype Ontology	genomic	425
GO	Gene Ontology	genomic, biological pathway	50713
HCPCS	Healthcare Common Procedure Coding System	clinical feature, drug sampling	7094
HGNC	HUGO Gene Nomenclature	genomic	32917
HP	Human Phenotype Ontology	biological feature	27482
ICDO	International Classification of Diseases Ontology	clinical feature	1313
LOINC	Logical Observation Identifier Names and Codes	clinical feature	268552
MONDO	Mondo Disease Ontology	clinical feature	43735
NCIT	National Cancer Institute Thesaurus	biological feature, clinical feature	166884
OAE	Ontology of Adverse Events	drug side effect, response to treatment	5762
OMIM	Online Mendelian Inheritance in Man	biological feature	97261
OPMI	Ontology of Precision Medicine and Investigation	clinical feature, clinical trial	2939
ORDO	Orphanet Rare Disease Ontology	clinical feature	14886
PDQ	Physician Data Query	clinical feature, drug screening	13452
PMAPP-PMO	PMO Precision Medicine Ontology	genomic, clinical feature, clinical trial, sampling	76154
SO	Sequence Ontology	genomic, transcriptomic	2707

clustering, as detailed in Section 3. The anchoring step also translated into 378 (lightweight) alignment runs resulting in a set of 378 pairwise alignments, as well as in an affinity matrix computed based on these alignments. The duration and total mappings found by the anchoring step are also presented in Table 3.

Figure 2 presents a heatmap representation of the semantic overlap computed by the anchoring step. Individual heatmaps for each component of the anchoring process are available as supplementary materials [5]. A few ontologies have a high number of direct cross-references between them or reuse classes from each other extensively. Logical definitions are less relevant to establish the semantic overlap between ontologies, since the majority of ontologies used does not declare them. The Lexical Matcher is the method that is able to find more correspondences for more ontology pairs.

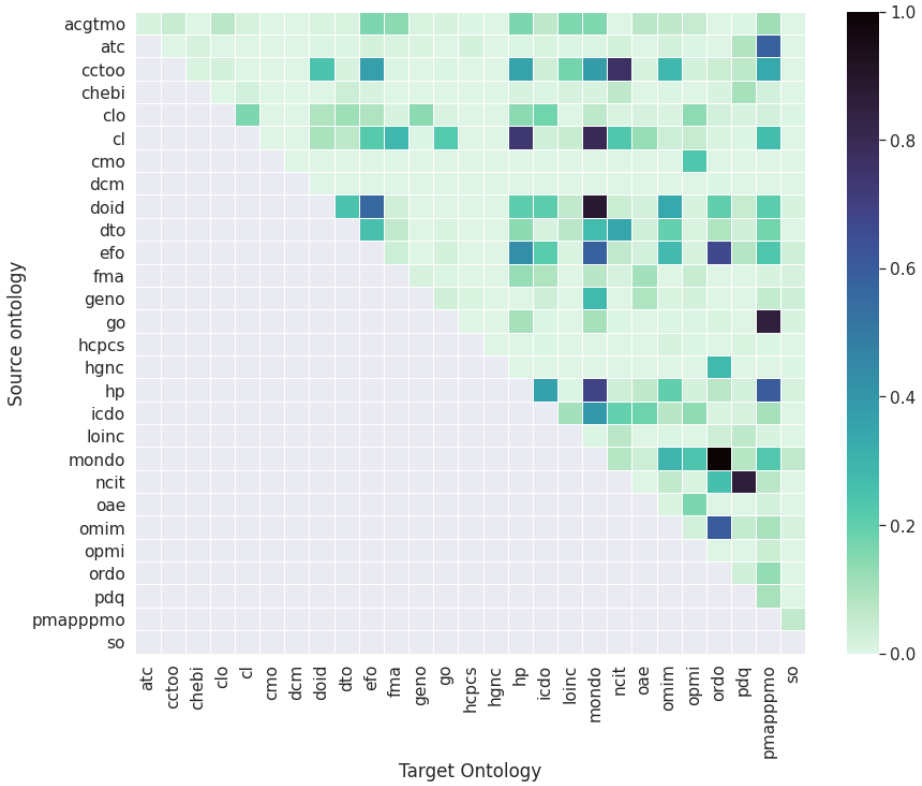

Fig. 2: Heatmap of the semantic overlap between ontologies

[5] https://github.com/liseda-lab/holistic-matching-aml

The affinity matrix was then used as input to clustering with spectral clustering. We tested cluster numbers between 3 and 6, and empirically selected 4 clusters which are shown in Table 2.

Table 2: Ontologies organized by cluster

Cluster	Ontologies	Classes
C1	NCIT, PDQ, LOINC, ChEBI, CCTOO	621079
C2	PMAPP-POM, GO, HP, ATC, FMA, CL, CLO, OAE, ACGT-MO, ICDO, SO, HCPCS, GENO	314820
C3	CMO, OPMI	5993
C4	MONDO, ORDO, DOID, OMIM, EFO, HGNC, DTO, DCM	249893

We applied the CPA and CIA strategies on the four clusters. In the CPA, the alignment tasks are run between all pairwise combinations of ontologies within each cluster, in the same manner as the GPA strategy. This translated into 117 pairwise alignment tasks and output alignments, which were merged within each cluster to produce 4 intermediate cluster KGs. In the CIA, only n-1 alignment tasks are required to integrate the n ontologies in each cluster incrementally, so 24 alignment tasks were necessary in total to produce 4 intermediate cluster KGs. The final step of each strategy was the merging of the cluster KGs through the anchoring algorithm to create a fully integrated KG. Again, the statistics of the alignment processes are summarized in Table 3.

Since we employed the GPA strategy only as a reference point for the state of the art, we did not perform the merging of the pairwise alignments into a single KG (as it would be beyond the state of the art). Thus, the GPA runtime is directly comparable to the sum of anchoring and within-cluster alignment runtimes for the CPA and CIA strategies. We note that, while the GPA takes more than 31 hours to complete, anchoring+CPA takes less than 24 hours and anchoring+CIA less than 16 hours. It is obvious that although matching times are greatly reduced in CPA (by nearly 12 hours) and CIA (by nearly 19 hours) when compared with GPA, the process of loading the ontologies using the OWL API is responsible a considerable portion of the time spent in running the alignment processes.

Table 4 presents the final alignment sizes produced by each strategy[6], where for CPA and CIA, the final alignment size results in combining the within-cluster mappings with between-cluster anchoring mappings, where the latter contributed over 200,000 mappings for both strategies. CPA produced a final KG that is 60% smaller than the total GPA mappings, while CIA produced one that is 65% smaller. We note that CPA strategy led to a greater number

[6] individual statistics available in the supplementary materials

Table 3: Alignment results

	Runtime (hh:mm)			Alignment	
Strategy	Load	Match	Total	Mappings	Tasks
GPA	11:47	19:51	31:37	554547	378
Anchoring	11:47	01:59	13:46	427300	378
CPA	02:25	07:42	10:07	219021	117
CIA	01:05	01:05	02:10	193503	24

GPA: global pairwise alignment. CPA: within-cluster pairwise alignment.
CIA: within cluster incremental alignment.

of mappings than the original global anchoring, whereas the CIA strategy led to a number of mappings just under the global anchoring. While this might suggest that the CIA strategy is losing relevant mappings, it is in fact a natural consequence of the incremental strategy, due to the fact that AML is configured to produce (mostly) 1 to 1 alignments. Thus, if in the pairwise strategy we have 3 mappings between equivalent classes c_1A, c_1B and c_1C of ontologies A, B and C, in the incremental strategy we would have only 2 mappings, since once ontologies A and B are combined into AB in a first iteration, AML will generally produce only 1 mapping between each class of AB and C, so c_1C would be mapped to either c_1A or c_1B, but not both. We note, however, that the third mapping would be semantically redundant, as it is implied by other two. Thus, the CIA strategy is expected to capture less mappings than the CPA strategy, but most of the missing mappings will be semantically redundant. A comparison of the alignments produced by CPA and CIA revealed that all CIA mappings are contained in the CPA alignment, with CIA covering 88% of the mappings found by CPA.

Table 4: Merged alignments results

Strategy	Total Mappings
GPA	554547
CPA+anchoring	442649
CIA+anchoring	417131

GPA: global pairwise alignment. CPA: within-cluster pairwise alignment.
CIA: within cluster incremental alignment.

4.4 Discussion

The holistic alignment of real world ontologies is a challenge that state of the art ontology matching systems that compete in the OAEI have yet to address. The very good performance of systems such as AML[17] and LogMap[14] in the biomedical tracks at OAEI[12] is impressive, but pales in comparison to the challenges of matching ontologies that have not stood the scrutiny applied to benchmark ontologies in organized challenges. In the course of this work, we encountered several hurdles due to syntactical issues in the ontologies or unexpected uses of some properties that had to be solved to ensure adequate coverage. As an example we highlight the case of the Experimental Factor Ontology (EFO) ontology that establishes cross-references between a single class and 77 classes in the Human Disease Ontology (DOID). A cross-reference is usually interpreted as an equivalent or closely related class, and this is explored by AML to produce equivalence mappings, but in this case the underlying relation between the one class and the 77 is one of subsumption. Addressing such cases correctly, will require adapting AML.

A recognized challenge in holistic matching is that the order of the matching tasks can impact the quality of the final alignment [13]. To circumvent this issue, [18]) developed a method that performs simultaneous matching of ontologies but unfortunately results in substantial losses in performance when compared to pairwise methods. The cost of determining the order for incremental matching is not considered by other works (e.g. [13]), however we argue that it must be considered a part of the alignment process. Moreover, employing simply lexical similarity is less than ideal in the biomedical domain where there is a high level of synonymy that is not always captured by the lexical component of the ontologies. In this work, we employ the same method to determine cluster affiliation and matching order, which is based on the semantic overlap between ontologies as measured by very high lexical similarity but also based on cross-references and logical definitions, which are particular to the biomedical domain. While in other works, clustering or tree mining is employed to determine the order of matching, we chose to apply clustering to actually partition the search space. This not only allowed a reduction of the matching tasks, but since clusters are based on semantic overlap and group together ontologies of the same domain, it can also mitigate the problem of false positives caused by homonyms. Let's take the example of the class *Gingiva* in the Foundational Model of Anatomy (FMA) ontology and the class *Gum* in the National Cancer Institute Thesaurus (NCIT). While 'gum' and 'gingiva' are synonymous words, in this case *Gum* actually refers to a type of chemical. However, since NCIT and FMA were actually placed in different clusters, the impact of these type of mappings can be minimized.

Although it is not possible to directly measure the quality of the resulting alignments short of a manual evaluation (as no reference alignments exist for these ontologies), an analysis of the number of mappings obtained can shed light on some interesting aspects. The GPA represents an upper bound on the number of mappings. It finds 120 thousand more mappings that anchoring, which we hypothesize to have a lower precision but increased recall, since the extra

method employed by the full AML pipeline compared to anchoring are mostly methods that were designed to increase recall, assuming that the performance of AML in these ontologies is comparable to its performance in the OAEI biomedical benchmarks. One advantage of the clustering-based approaches is that they have the potential to increase the precision of mappings between clusters, by only establishing mappings based on the high precision and lower recall anchoring strategy, while increasing recall within the clusters, by employing more sophisticated alignment methods. Moreover, as detailed in Section 4.3, the CIA strategy is expected to find less mappings than the CPA strategy, but these will be mostly semantically redundant mappings.

5 Conclusions

The rich panorama of both publicly available data and ontologies in the biomedical domain represents an opportunity for developing explainable knowledge-enabled systems. In multi-domain areas, such as personalized medicine, this requires the integration of multiple data sources and ontologies. Holistic ontology matching and integration holds the promise to scale semantic data integration to multiple sources [28], however holistic ontology matching in the biomedical domain is still an open challenge.

We have developed a novel approach for holistic ontology matching that builds on an existing system, AML[9], addressing the requirements and challenges of the biomedical domain. We demonstrated that the straightforward application of the pairwise alignment approach to all ontology pairs takes up to 100% more time than the novel clustering-based approaches. We further demonstrated that the within-cluster incremental alignment approach is five times faster than the within-cluster pairwise alignment approach. All approaches were able to generate a fully integrated KG, meaning that all ontologies have mappings to one or more of the other ontologies, effectively responding to the coverage requirement. The quality assessment of the resulting alignment is not straightforward, since there are no holistic reference alignments within the biomedical domains, and out of the 378 pairwise alignments, only one pair is covered by an existing reference (FMA-NCI) but it employs an outdated version of the ontologies and was produced semi-automatically.

The proposed approach can be extended with further refinements. To increase the coverage and semantic richness of the KG, complex mappings can be applied to more accurately capture the relations between their entities. While the KG construction will be mostly automated, expert feedback will be paramount to ensure an accurate KG that can support explanations. To make the most efficient use of feedback, we will develop algorithms to identify potentially doubtful mappings that require user validation, and algorithms that propagate the user feedback automatically [6].

The experience of applying a state of the art ontology matching system to a large set of real world biomedical ontologies for holistic matching and integration resulted in lessons learnt for future endeavours. One of the identified challenges

was the comparative evaluation of the alignment quality produced by pairwise and holistic approaches. One future opportunity is to build upon the set of reference alignments made available by the OAEI to create a holistic reference alignment following the approach described by Roussille *et al.*[29]. Another lesson was the fact that the ontology loading times slow down the alignment process substantially, also this was partly due to the fact that AML still employs an older version (3.4) of the OWL API. Preliminary testing showed that a new version of the OWL API (5.1) speeds up the loading by a factor of 2. Perhaps the biggest challenge was in handling the varying degrees of quality of the ontologies, with formatting issues and non-standard uses of the cross-reference property that required *ad hoc* solutions to circumvent, and will require further extensions to AML to handle more adequately.

Building on decades of work by the semantic web and biomedical ontologies communities, we have developed an approach for holistic matching and integration of ontologies from multiple domains to build KG to support AI-based personalized cancer therapy. The size, diversity and complexity of the underlying ontologies and overarching domain represented significant challenges that required evolving the current state of the art in ontology matching.

Acknowledgments This work was supported by FCT through the LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020). It was also partially supported by the KATY project which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101017453.

References

1. Babalou, S., Grygorova, E., König-Ries, B.: Comerger: A customizable online tool for building a consistent quality-assured merged ontology. In: European Semantic Web Conference. pp. 19–24. Springer (2020)
2. Caldarola, E.G., Rinaldi, A.M.: An approach to ontology integration for ontology reuse. In: 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI). pp. 384–393. IEEE (2016)
3. Chari, S., Gruen, D.M., Seneviratne, O., McGuinness, D.L.: Directions for explainable knowledge-enabled systems. arXiv preprint arXiv:2003.07523 (2020)
4. Chari, S., Gruen, D.M., Seneviratne, O., McGuinness, D.L.: Foundations of explainable knowledge-enabled systems. arXiv preprint arXiv:2003.07520 (2020)
5. Chatterjee, N., Kaushik, N., Gupta, D., Bhatia, R.: Ontology merging: A practical perspective. In: International Conference on Information and Communication Technology for Intelligent Systems. pp. 136–145. Springer (2017)
6. Cruz, I.F., Stroe, C., Palmonari, M.: Interactive user feedback in ontology matching using signature vectors. In: ICDE 2012. pp. 1321–1324. IEEE (2012)
7. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag, Heidelberg (DE), 2nd edn. (2013)

8. Faria, D., Pesquita, C., Mott, I., Martins, C., Couto, F.M., Cruz, I.F.: Tackling the challenges of matching biomedical ontologies. Journal of biomedical semantics **9**(1), 1–19 (2018)
9. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M.: The agreementmakerlight ontology matching system. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". pp. 527–541. Springer (2013)
10. Ferreira, J.D., Teixeira, D.C., Pesquita, C.: Biomedical ontologies: Coverage, access and use. In: Wolkenhauer, O. (ed.) Systems Medicine Integrative, Qualitative and Computational Approaches, pp. 382 – 395. Academic Press, Elsevier (2020). https://doi.org/https://doi.org/10.1016/B978-0-12-801238-3.11664-2, http://www.sciencedirect.com/science/article/pii/B9780128012383116642
11. Gruetze, T., Böhm, C., Naumann, F.: Holistic and scalable ontology alignment for linked open data. LDOW **937**, 1–10 (2012)
12. Harrow, I., Jiménez-Ruiz, E., Splendiani, A., Romacker, M., Woollard, P., Markel, S., Alam-Faruque, Y., Koch, M., Malone, J., Waaler, A.: Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. Journal of biomedical semantics **8**(1), 1–13 (2017)
13. Hertling, S., Paulheim, H.: Order matters: Matching multiple knowledge graphs. arXiv preprint arXiv:2111.02239 (2021)
14. Jiménez-Ruiz, E.: Logmap family participation in the oaei 2020. In: Proceedings of the 15th International Workshop on Ontology Matching (OM 2020). vol. 2788, pp. 201–203. CEUR-WS (2020)
15. Köhler, S., Bauer, S., Mungall, C.J., Carletti, G.O.N., Smith, C.L., Schofield, P.N., Gkoutos, G.V., Robinson, P.N.: Improving ontologies by automatic reasoning and evaluation of logical definitions. BMC Bioinformatics **12**, 418 – 418 (2011)
16. Lecue, F.: On the role of knowledge graphs in explainable AI. Semantic Web **11**(1), 41–51 (2020)
17. Lima, B., Faria, D., Couto, F.M., Cruz, I.F., Pesquita, C.: Oaei 2020 results for aml and amlc. (2020)
18. Megdiche, I., Teste, O., Trojahn, C.: An extensible linear approach for holistic ontology matching. In: International Semantic Web Conference. pp. 393–410. Springer (2016)
19. Noy, N.F., Shah, N.H., Whetzel, P.L., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids res **37**(2), W170–W173 (2009)
20. Oliveira, D., Pesquita, C.: Improving the interoperability of biomedical ontologies with compound alignments. Journal of biomedical semantics **9**(1), 1–13 (2018)
21. Osman, I., Ben Yahia, S., Diallo, G.: Ontology integration: Approaches and challenging issues. Information Fusion **71**, 38–63 (Jul 2021)
22. Otero-Cerdeira, L., Rodríguez-Martínez, F.J., Gómez-Rodríguez, A.: Ontology matching: A literature review. Expert Systems with Applications **42**(2), 949–971 (2015)
23. Pesquita, C.: Towards semantic integration for explainable artificial intelligence in the biomedical domain. In: BIOSTEC 2021. vol. 5, pp. 747–753 (2020)
24. Pesquita, C., Faria, D., Santos, E., Couto, F.M.: To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. Ontology Matching (2013)
25. Pesquita, C., Faria, D., Stroe, C., Santos, E., Cruz, I.F., Couto, F.M.: What's in a 'nym'? synonyms in biomedical ontology matching. In: International Semantic Web Conference. pp. 526–541. Springer (2013)

26. Pour, N., Algergawy, A., Amini, R., Faria, D., et al.: Results of the ontology alignment evaluation initiative 2020. In: OM 2020. vol. 2788, pp. 92–138. CEUR-WS (2020)
27. Rahm, E.: Towards large-scale schema and ontology matching. In: Schema matching and mapping, pp. 3–27. Springer (2011)
28. Rahm, E.: The case for holistic data integration. In: East European Conference on Advances in Databases and Information Systems. pp. 11–27. Springer (2016)
29. Roussille, P., Megdiche, I., Teste, O., Trojahn, C.: Boosting holistic ontology matching: Generating graph clique-based relaxed reference alignments for holistic evaluation. In: European Knowledge Acquisition Workshop. pp. 355–369. Springer (2018)
30. Saleem, K., Bellahsene, Z., Hunt, E.: Porsche: Performance oriented schema mediation. Information Systems **33**(7-8), 637–657 (2008)
31. Silva, M.C., Faria, D., Pesquita, C.: Integrating knowledge graphs for explainable artificial intelligence in biomedicine? In: Ontology Matching workshop at the International Semantic Web Conference (2021)
32. Stoilos, G., Geleta, D., Shamdasani, J., Khodadadi, M.: A novel approach and practical algorithms for ontology integration. In: International Semantic Web Conference. pp. 458–476. Springer (2018)