Measuring the quality of ontology matches in BioPortal

Tobias Räwer, Lina Molinas Comet, and Jonas Rülfing

RWTH Aachen University, Aachen, Germany
{tobias.raewer,lina.molinas.comet,jonas.ruelfing}@rwth-aachen.de
http://dbis.rwth-aachen.de/cms

Abstract. BioPortal [1] is a repository of biomedical ontologies which contains not only the ontologies as isolated parts but also matches or mappings between them. Those mappings are necessary because more than one ontology can refer to the same domain and describe similar entities and relations. However, partial overlaps between ontologies can exist, as a consequence of linking their entities. In this paper we explain our approach of measuring the quality of the matches between two ontologies in BioPortal. In this sense, we describe the metrics we used in order to evaluate the suitability of the existing ontology matches, the comparison of results, and our findings when contrasting our results against the silver standard from Pistoia [2]. We consider that being able to detect wrong matches will be beneficial in order to improve the general quality of the available mappings.

Keywords: ontology matching · similarity measures · ontology evaluation · BioPortal

1 Introduction

Ontologies are helpful for describing and formalizing concepts, and with this to share a common understanding of an entity or relation [1]. However, when more than one ontology describe the same concepts, inconsistencies may arise and the common interpretation is lost. Fortunately, methods for ontology mapping were develop to ensure that two ontologies can be merge or align together and also preserve their properties and restrictions [2]. Despite the fact that ontology mappings represent an advance in the unification of concepts, there are still problems when creating those mappings, which can result in wrong matches. BioPortal is an environment in which biomedical ontologies and mappings among them are stored. However, there is not information about the quality measures applied to those matches to conclude if they are good. With this in mind, our contribution in this paper is the evaluation of the existing matches in BioPortal by using similarity metrics including structural and lexical similarity measures, and by

[1] http://BioPortal.bioontology.org
[2] https://www.pistoiaalliance.org

calculating scores to determine how suitable a match is. The structure of this paper is as follows: first we describe the data source and our solution proposal. Then, in section 2 we introduce some concepts in connection to the topic. In section 3 we present well-known approaches in ontology matching, the common problems related to it, and the techniques for ontology matching evaluation. After that, in section 4 we describe our data selection criteria, and the different similarity metrics used in this project. Then, in section 5 we present the results of our implementation, and the corresponding evaluation. Finally, in section 6 we present our conclusions and reflections about the future work to be done.

1.1 Data Source Description

BioPortal is an open repository of biomedical ontologies [3]. To this day, there are 792 ontologies and a high number of mappings among them. The BioPortal provides a service to access the ontologies (the NCBO Web service REST API), and a set of tools to work with them, namely search tools, to access data such as mappings or biomedical resources [4].

1.2 Problem statement

In the BioPortal environment there are a high number of entities which are matched with each other, which is helpful for knowledge discovery. Anyhow, it may also be detrimental, if the matches among the entities are not correct. For this reason, it is important to guarantee the quality of those matches. However, in the mentioned environment there are not publicly available statistics or measures indicating the quality of the ontology matches among entities in BioPortal.

1.3 Solution Proposal

Based on our problem statement in 1.2, our proposal is the evaluation of the suitability of the existent matches in order to determine how good is one match and how it affects the overall mapping among entities. Moreover, we are also interested in detecting loops in the existing matches. To reach our goal, we first applied well-know similarity measures, as well as other metrics. After that, we evaluated our results by testing our findings against the silver standard from Pistoia, and finally we calculated precision and recall values to reach our conclusions.

2 Background

For a better understanding, in this section we define the two main concepts related to the topic under study. After that, we present previous works related to ontology matching and its evaluation.

2.1 Terminology and Definitions

Ontology An ontology is a specification of a conceptualization, i.e. is the description of the concepts and existent relations for an individual or group of individuals or entities [5]. More precisely, it is an explicit and formal description of the concepts in a given domain, the properties or roles of those concepts, and the existent restrictions (if they are), for example in the roles among the concepts [1]. Moreover, ontologies usually follow a hierarchy when describing the involved concepts [2]. Ontologies can be used for different purposes, for instance: to share a common understanding of the structure of the information among people and/or software agents, to reuse domain knowledge, and to make the implicit assumptions of a domain explicit[1].

Ontology Mapping An ontology mapping, based on the definition of Kalfoglou and Schorlemmer [2], is the existent relation between two ontologies and where the concepts, roles, restrictions are preserved. In a formal and more theoretical aspect, there exist total mappings between ontologies where all interpretations of one ontology match the other. But there are also partial ontology mappings where only a subset of the interpretations is matched [2].

The relevance of ontology matching lies in the fact that it allows the reuse of the ontologies. Moreover, it enables the finding of connections among concepts in different ontologies, and with this, the extraction or reuse of the knowledge contained in the related ontology [6].

3 Related Work

In this section, we present a brief summary of the common approaches to tackle the ontology matching, as well as the current underlying problems. After that, we introduce some techniques for ontology matching evaluation.

3.1 Common approaches in ontology matching

Based on the survey presented by Hooi et al. [7], and with strong reference to the work of Euzenat and Shvaiko [8], there are three main categories of ontology matching: terminological mapping, structural mapping, and semantic mapping. In the *terminological mapping*, which uses token analysis to reduce a word to a common format and establishing importance of the term based on weighting of relations by comparison of paths, the following techniques are highlighted:

- **String-based:** it takes into account the edit distance (Levenshtein distance) by counting and normalizing the required editing operations to transform the first word into the second. Example: bag of words, n-grams, exact matching, tokenization, etc.
- **Language-based:** it considers the punctuation and cases to tokenize the strings, then uses lemmatization to find possible basic forms of the word. It uses NLP techniques such as lexicon.

- **Linguistic resource:** it uses sources such as WordNet for linguistic knowledge to interpret strings (synonyms, hyponym, antonym, etc.)

In the *structural mapping* category, which looks at the relationship between concepts in the ontology structure, the techniques are the following:

- **Taxonomy mapping:** it uses the concept of super-concept rule and bounded-paths. Super-concept indicates that if two concepts share the same parent concept, then it is assumed that those concepts are similar. On the other hand, bounded-paths compares two paths to identify similar concepts along that path.
- **Tree-based mapping:** it takes into account the position of the nodes on the graph. Neighbour nodes are assumed to be similar if two nodes from two ontologies are similar.

On the *semantic mapping* side, which uses model of theoretic semantics to define well-formed-formula to express the meaning of anything without ambiguity, the techniques are the following:

- **External ontologies:** an intermediate reference ontology is used to provide general concepts and axioms, and clarify the meaning of the concepts and relations on a domain. This reference ontology can be an external ontology (WordNet, GFO, DOLCE, etc.), or one ontology that is built on the fly by using lexicons.
- **Deductive techniques:** it applies subsumption relation by satisfiability tests to check the existing mappings and discard poor mappings. This technique is apply in the merge of two ontologies.

In the context of *satisfiability* of mappings, the common techniques are: propositional satisfiability (SAT), modal satisfiability, and description logic (DL).

3.2 Current problems in ontology matching

As previously mentioned, ontology matching may be challenging and problems may occur when matching them. Based on Heflin and Song [9], the main problems in ontology matching are summarized in Table 1 In Table 2 we present some of the most general problems when matching ontologies. More precisely, the lack of an undefined ontology is due to the fact that each ontology represents an own perspective of a domain, and therefore it can exist a partial overlapping among ontologies. Even if the ontologies are highly similar, there is still the fact that they can be represented in different languages (multilinguism) [6].

Due to existing ontology matching problems, some methods were developed to minimize them. One of them include taking into account the different semantic relations among ontologies. Those relations can belong to different categories, for example: the relation among individuals (*owl:sameAs, owl:differentFrom, owl:inverseFunctionalProperty*, etc.), the relation among classes (*owl: equivalentClass, rdfs:subClassOf, owl:disjointWith*, etc.), the relation individual-class, the relation property-property (*rdf:subPropertyOf, owl:equivalentProperty, owl:disjointProperty*, etc.) [6].

Technique	Problems
String matching	- If similarity matching between labels of two (or more) ontologies is computed and the highest similarity is taken into account, then it could be that the results bring many false positives (e.g. entities referring to John Smith but without real relation between entities besides the entity name)
String matching combined with logical reasoning	- Reasoning based approaches highly depend on the correct expressions of the ontologies (e.g. the surname of a person may change after marriage). - Could not work in environments without formal semantics
Blocking: subdividing entities into mutually exclusive blocks (only those within the same block are compared)	- The required expertise may not be available (in case of manually subdivision) - Rules are not defined (then there is not ontology to contrast to) - When supervised learning techniques are applied to learn a blocking scheme, it could be that there is not enough training data

Table 1: Current problems in ontology matching by technique

General issues in ontology matching
- Large scale evaluation
- Performance of ontology matching techniques
- Challenges in discovering missing background knowledge
- Uncertainty in ontology matching
- User involvement - Explanation of matching results
- Multilingualism

Table 2: General problems in ontology matching [10]

3.3 Current techniques for evaluation of ontology matching

In the area of ontology matching, there is a lot of ongoing research. To streamline the efforts and to provide a platform for evaluation, comparison and discussion, the Ontology Alignment Evaluation Initiative (OAEI)[3] was founded in 2004. They provide different ontology mapping challenges every year and two different evaluation platforms for these, HOBBIT[4], based on the linking benchmark SPIM-BENCH [11], and SEALS[5].

Provided a mapping between two ontologies generated by an algorithm, the basic evaluation approach is the comparison of each individual link against a given reference alignment which was created by a group of domain experts.

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

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

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

With the results of this comparison, following commonly used metrics [12] are calculated. Abbreviations used are true positives(TP), false positives (FP), true negatives (FN) and false negatives (FN).

- **Precision** $= \frac{TP}{TP+FP}$; precision measures the correctness
- **Recall** $= \frac{TP}{TP+FN}$; recall measures the completeness
- **Fall-out** $= \frac{FP}{TN+FP}$; fall-out(X) measures the expectancy of false positives
- **F1 measure** $= \frac{\mathbf{2*precision*recall}}{\mathbf{precision+recall}}$
  the F1 measure(X) puts precision and recall into a relationship
- **Runtime, runtime complexity, scalability**: Next to the results, it is also important to rate the performance of an algorithm. Ontologies in practice can be huge, so it is important that the mappings can be generated in a feasible time.

These metrics are used by the OAEI to compare the results achieved by different algorithms. They are good indicators to see which ones are performing better on different sets of ontologies. The results are used to create a dialog and to continuous improve the existing algorithms.

Because of time constraints, we were not able to test our solution on some of the challenges with the provided evaluation platforms, but an evaluation, probably with HOBBIT because it is the more modern approach, would be one of the next steps. Nevertheless, the same metrics used by the OAEI were used to compare our results against the available reference.

4 Material and Methods

In this section, we describe our selection criteria when choosing the ontologies to evaluate, as well as a description of the standard against which we tested our proposed solution. Moreover, we explain in context the different metrics we implemented and used to evaluate the ontology mappings.

4.1 Data Selection Criteria

In order to evaluate the suitability of the matches in BioPortal, our selection criteria was to rely on the information provided regarding the number of mappings among the ontologies, and select those ontologies which are highly similar to each other. In this context, we included the ontologies Human Phenotype Ontology (HP), and Mammalian Phenotype Ontology (MP) to test our evaluation against the provided mappings from Pistoia.

It is worth to mention that we selected these ontologies as a sample for our evaluation, and based on the criteria, we considered that they represent a good candidate of the ontologies and mappings in BioPortal.

As mentioned before, we also take into consideration the silver mappings from Pistoia, which represent a benchmark of good and bad mappings in BioPortal.

4.2 Metrics for Quality Measurement of Matches

As already listed in section 3, there are several approaches to determine the quality of a match between two classes of an ontology. In this section we will present the metrics we used in our proposal to measure the quality of a link between two classes of two ontologies.

Jaccard, Sigmoid, and Cosine Similarity The Jaccard Similarity, or Jaccard Index, is one of the most well-known metrics which is used to determine the similarity between two sets, in this case the two ontologies. And it is given by the formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

Where A and B represent the two compared objects, and where the shared components (intersection) between the two objects are divided by the distinctive features (union minus intersection).

The maximum value in the Jaccard index is 1, indicating the highest similarity between the two compared entities. While 0 indicates that no similarity exist among the compared objects.

Something to remark is that even if this metric is suitable for the purpose of determining similarity, it is really sensitive or unstable when the size of the samples is small. [13]

Another metric which we considered is the Sigmoid Similarity, and it was introduced in [14] as an improvement of the Jaccard Similarity. Its formula is given as:

$$SigmoidSimilarity(A,B) = \frac{(e^{CF(A,B)} - 1)}{(e^{CF(A,B)} + 1)(DF(A) + DF(B) + 1)} \tag{2}$$

Where A and B represent the compared objects (similar to the Jaccard Similarity). Moreover, CF indicates the common features, and DF indicates the distinctive features. The extreme values 1 and 0 also have the same interpretation as in the Jaccard Index.

$$CosineSimilarity = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}} \tag{3}$$

Based on [15], we also implemented the Cosine Similarity metric. This also popular metric compares and measures the similarities between two non-zero vectors, in the context of Bag of Words (BOW).

Jaccard and Cosine Similarity applied to labels, definitions, and synonyms In order to determine the similarity between two nodes of the ontologies, our first approach was to compare some properties of the classes of the two ontologies. For that, we applied the three metrics previously described. We compared the preferred labels of the classes of the two ontologies, as well as the definitions of the classes, and the synonyms. When calculating the Jaccard and Sigmoid similarity measures, we got the labels of each class in the two ontologies and compare them against each other. We proceeded in the same manner with the definitions. However, when comparing the synonyms, we got the labels of one class from ontology A and compared them against the synonyms of the other class in ontology B.

When calculating the Cosine Similarity, first we created two vectors one containing the label of the class from ontology A, and the other the label from ontology B respectively. Then, the Cosine Similarity is determined by comparing the frequency of the words in both of them, as well as the differences. Two vectors are maximally similar if the value calculated for the cosine similarity is 1, and are dissimilar if the value is 0. We proceeded in the same manner when comparing the definitions and synonyms of the classes.

Jaccard and Cosine Similarity applied to subtrees The aforementioned implementations just consider the similarities of two nodes. To also be able to measure how similar the subgraphs of two nodes are, we applied the Jaccard and Consine Similarity to the respective subgraphs in the following manner:

Given two ontologies O_1 and O_2, a mapping $M \subseteq V_{O_1} \times V_{O_2}$, $O_1 \cup O_2$, two nodes $a \in O_1$ and $b \in O_2$ and the subgraphs A and B where A is the subgraph of a and B the subgraph of b. Then we define $A \cap B := \{x | x \in V_A \wedge \exists x' \in V_B : (x, x') \in M\}$ and $A \cup B = A + B - A \cap B$. Then we can calculate the Jaccard Similarity for the two subtrees accordingly as $\frac{A \cap B}{A \cup B}$.

For our implementation we considered that the subtree excludes the root nodes - i.e. a and b - because the metric should give a reason about the subgraphs and not the nodes itself. This was done accordingly for the Cosine Similarity.

For the cosine similarity, the label of every node of each subgraph was split into a set of words, and then all sets joined to one set for each subgraph. Then the similarity between these two sets was measured with the cosine similarity.

Children Metric The *Children Metric* is a metric, which focus on the structure of the ontologies. In Figure 1 the basic idea of the metric is illustrated and was originally motivated by Pathak and Chute [16]. For this metric we assume that the mappings are valid and that the ontologies are consistent. This means, whenever one class of an ontology matches to another class in a different ontology, sub-classes should also correlate to each other. Figure 1 shows us two ontologies and matches between A and B in O_1 and C and D in O_2. B is a subClassOf A in O_1 and C and D are disjoint in O_2. By our assumption, we can imply that D should be a subClassOf C, however in this example they are disjoint and therefore we have a logical inconsistency. This implies that either m_1 or m_2 is a bad match. Our *Children Metric* is able to detect these inconsistencies.

Loop detection When merging two ontologies, logical inconsistencies may arise. Given two ontologies O_1 and O_2 and a mapping $M \subseteq V_{O_1} \times V_{O_2}$, $O_1 \cup O_2$

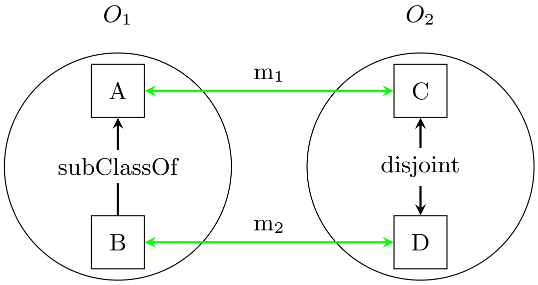


Fig. 1: Example of an inconsistent mapping detected by the *Children Metric*. Adopted from [16]

might contain loops. These loops are logical errors, because they imply that nodes exist which are ancestors and descendants of themselves. In figure 2 an example of such a case can be seen in an abstract representation.

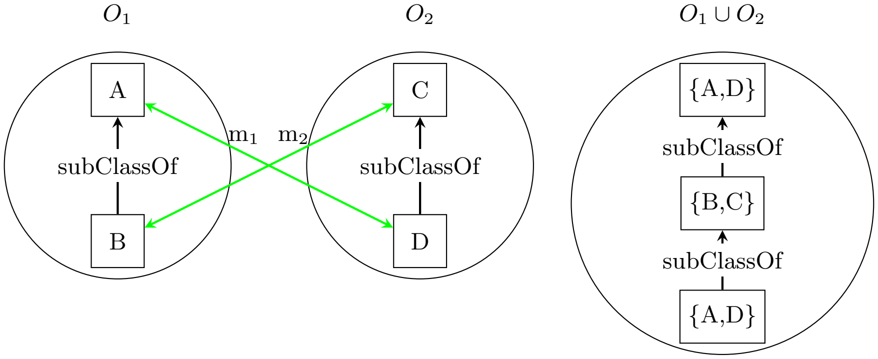


Fig. 2: Example of a loop induced by the two matches m_1 and m_2 during a merge

In different papers as [17] and [19], loops where identified as problems, because they are logical errors which imply that there is an error in either one of the two ontologies O_1 or O_2 or there is at least one wrong match $m \in M$. Additionally, existing loops can break algorithms which are supposed to run on loop free ontologies [18].

These loops are relevant for this paper, because they indicate that there might be an error in one of the matches. For time constraint reasons, only simple loops between two ontologies induced by exactly two matches are analyzed. In a future work this could be extended.

The implemented loop check analyzes every match and checks if there is a second match which would induce a loop.

Given a match $(a, b) \in M$, it is checked if there exists a match between an ancestor of a and a descendant of b and vice versa. If this is the case, there exists a loop in $O_1 \cup O_2$.

5 Experiment

In order to evaluate the quality of existing mappings in BioPortal, we implemented the metrics presented in the previous section and compared the results against the silver standard from Pistoia.

5.1 Implementation

We implemented our approach in Java, since it provides useful libraries to work with ontologies, such as *Apache Jena*[6]. The workflow of our implementation is illustrated in figure 3. First the user needs to provide two ontologies O_1 and O_2 in *.rdf* or *.owl* format, which are read in and represented as an instance of the OntModel class provided by *Apache Jena*. If the user also provides a mapping file, which is a text file with two columns, where each column represents the URI of a class in O_1 and O_2 respectively, the file is also read in. If no such file is present, we fetch the mappings from the BioPortal, using the REST API[7]. When we gathered all the data, we create an instance of EvaluationModel. The EvaluationModel class provides the necessary datastructures to store all the calculated metric scores for each match. After that, we calculate for each match the scores for the different metrics presented in the previous section.

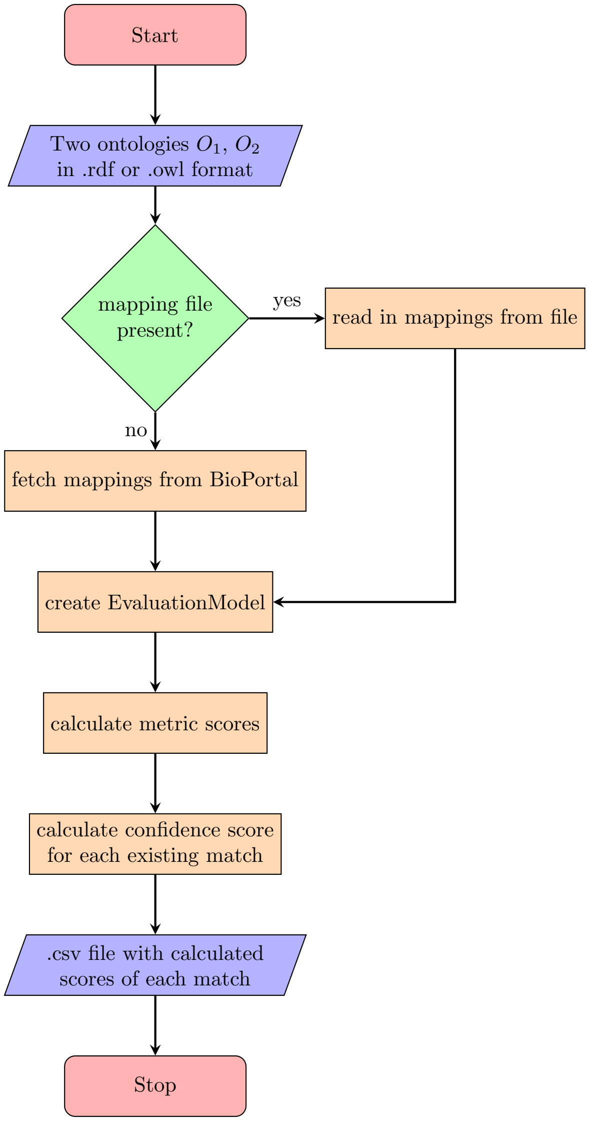

Start
Two ontologies O_1, O_2 in .rdf or .owl format
mapping file present?
yes
read in mappings from file
no
fetch mappings from BioPortal
create EvaluationModel
calculate metric scores
calculate confidence score for each existing match
.csv file with calculated scores of each match
Stop

Fig. 3: Flowchart of the program flow of our implementation

Based on the results of the metric scores we calculate a confidence score between 2 and 4, similar to the silver standard. The confidence score is based on

[6] https://jena.apache.org/
[7] http://data.bioontology.org/documentation

a "naive" decision tree, which is illustrated in figure 4. The decision tree uses "intuitive" rules to calculate a score. First, we check if the mapping contains a loop. If this is the case, the probability is high that the match is bad, thus we assing in this case the score 2. If it does not contain a loop, we check the result of the *Children Metric*. If the result is true, we take the minimum of the calculated similarity metrics for the labels and compare if it is under the threshold of 0.8. If it is, we set the score to 2, otherwise to 3. In the case, where the result of the *Children Metric* is false, we do the same but set the score to 3 if the result is under 0.8 and to 4 otherwise. Finally, we output a *.csv* file, which contains all results we calculated. For each match we output the URIs of the classes in O_1 and O_2, and the calculated metric results and our calculated score. The implementation and details can be found in the git repository (https://git.rwth-aachen.de/tobias.raewer/kglab-project).

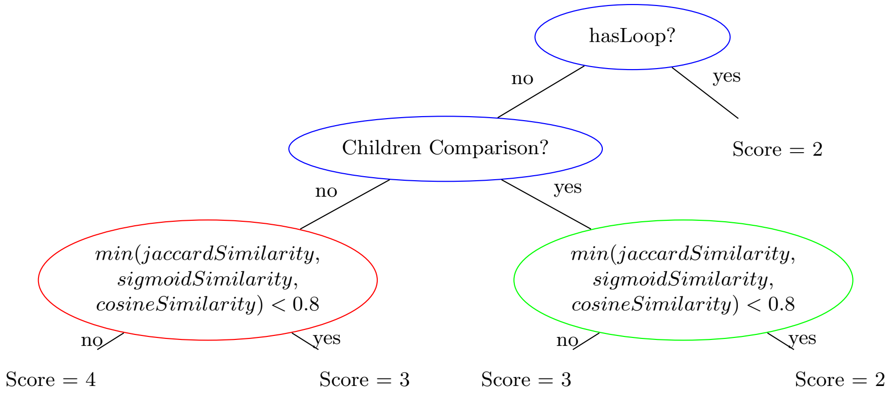

hasLoop?
no
yes
Score = 2
Children Comparison?
no
yes
$min(jaccardSimilarity, sigmoidSimilarity, cosineSimilarity) < 0.8$
$min(jaccardSimilarity, sigmoidSimilarity, cosineSimilarity) < 0.8$
no
yes
no
yes
Score = 4
Score = 3
Score = 3
Score = 2

Fig. 4: Intuitive decision tree to calculate the confidence score of the quality of a match

Ontology	Acronym	# Classes
Human Phenotype Ontology	HP	18,154
Mammalian Phenotype Ontology	MP	14,028

Table 3: Ontologies used for evaluation

5.2 Evaluation

To evaluate our results, we use the *.csv* file given as output of our program. We compared our results against the provided scores by the silver standard of Pistoia. Since they provided the scores for mappings between the Human

Phenotype Ontology (HP) and the Mammalian Phenotype Ontology (MP), we also used these two ontologies in our evaluation (cf. Table 3). We downloaded the latest version of both ontologies in *.xrdf* format from the BioPortal website [8]. To retrieve our results, we ran our implementation twice. In the first run, we only used the mappings we retrieved from BioPortal using the REST API. In the second run, we provided a mapping file, which contains all the mappings, we have a score from the silver standard (cf. Table 4).

Ontology Pair	**Retrieved Mappings (REST API)**	**Mappings (Silver Standard)**
HP-MP	750	2005

Table 4: Number of retrieved mappings

After obtaining the results of both runs, we used a Jupyter Notebook to also compare our ”naive” score against machine learning techniques. We created two models, to predict the score based on our metric scores. One which uses a multinomial logistic regression estimator, and one which uses a random forest estimator. Therefore, we used the silver standard as ground truth and split the data into a test set and a training set. The test set contains 40% of the data. General statistics of the sets can be found in Table 5.

Approach	**#Mappings**	**Scores (silver standard)**		
		2	**3**	**4**
Naive (REST API)	557	28	7	522
Logistic regression (REST API)	223	12	4	207
Random forest (REST API)	223	12	4	207
Naive (Mapping file)	2005	406	248	1351
Logistic regression (Mapping file)	802	174	93	535
Random forest (Mapping file)	802	174	93	535

Table 5: Statistics of test sets used for evaluation

The results after estimating the scores with the machine learning techniques and comparing our ”naive” scores against the silver mapping scores are shown in Table 6.

[8] https://BioPortal.bioontology.org/

Approach	Precision	Recall	F1-measure
Naive (REST API)	0.88389	0.88509	0.88352
Logistic regression (REST API)	0.86165	0.92825	0.89371
Random forest (REST API)	0.90198	0.91479	0.90835
Naive (Mapping file)	0.66276	0.49027	0.52989
Logistic regression (Mapping file)	0.61291	0.69576	0.62828
Random forest (Mapping file)	0.66611	0.71321	0.66593

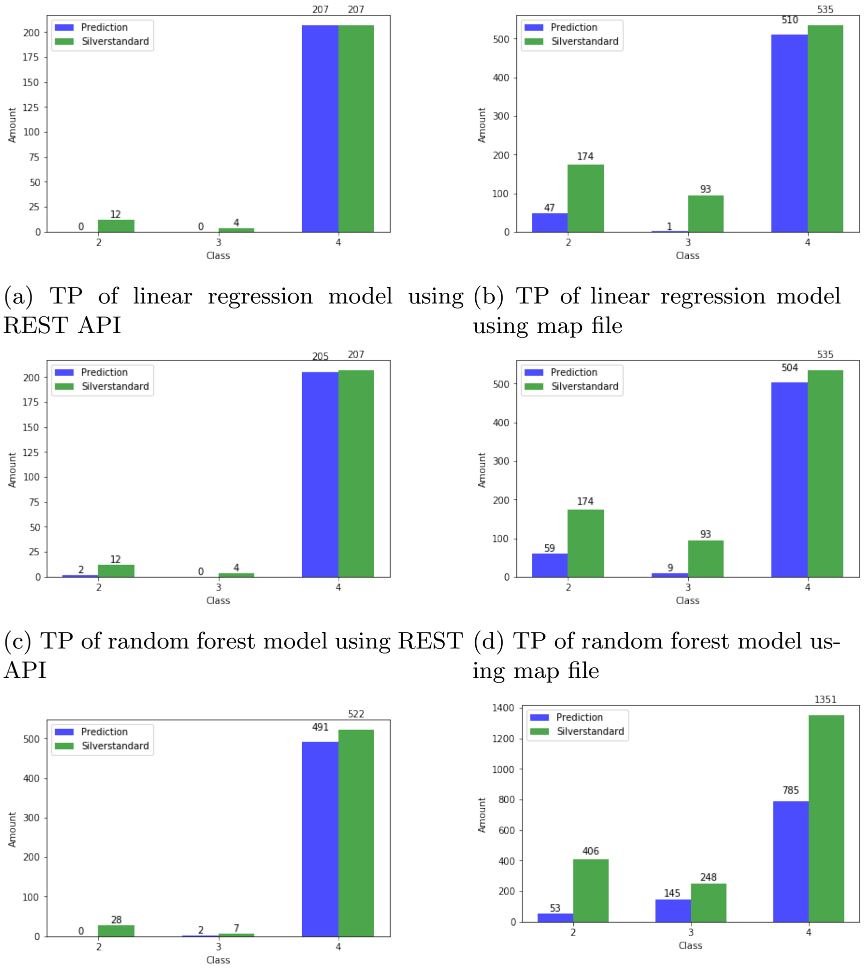
Table 6: Results of our evaluation

As one can see, our predictions of the scores for quality of a match using the mappings obtained by the REST API seem to be very good. All three approaches reach nearly a F1-measure of 90%, where the random forest approach seems to be the best approach with a precision of 90.198% and a F1-score of 90.835%. However, if we take a look at Figure 5, we see that these numbers have to be interpreted very carefully. The data set we are using for our evaluation using the REST API contains mostly matching with the score 4. We only have 28 matches with score 2 and 7 matches with score 3. This is not enough data to train the machine learning estimators sufficiently. As one can see, the linear regression model predicts all matches with score 4 correctly, but none with the score 2 or 3. Nevertheless, it reaches a F1-measure of nearly 90%.

Considering the mappings we measured using the mapping file, we see that the scores decrease. Our naive approach reaches only an F1-measure of 52.989%. We predicted 58.11% of the score 4, 58.47% of score 3, and 13.05% correctly. With the machine learning approaches the results improve, where the random forest approach gives us the best results. We reach a precision of 66.611% and a F1-measure of 66.593%. Furthermore, we see that the predictions for score 4 are very good with 94.21% (cf. Figure 5). However, the predictions for score 3 are only 9.677%, which is not very good. Nevertheless, it is important to mention that also this test set is relatively small and the scores in the silver standard are provided by domain experts. They might know things about matchings, which our approach currently does not consider and therefore it is hard to estimate the quality of a match.

6 Conclusion and Future Work

In this project we have developed a framework, for measuring the quality of existing mappings between two ontologies. Our approach considers several metrics to assess the quality based on lexical similarity and structural similarities and finally give an estimation how good the quality of a match is by providing a score. This score is calculated by an “naive” approach and two different machine learning techniques. Then, we have performed a evaluation of our proposal, to see how our metrics perform. Therefore, we evaluated the mapping quality between the Human Phenotype Ontology and the Mammalian Phenotype Ontology, and used the silver standard provided by Pistoia as reference data and ground truth. Our results seem to be very good, if we consider mostly matches based on lexical similarity (e.g. found by the LOOM algorithm). In this case we get for all three approaches a F1-measure of nearly 90%. However, if we increase to number of mappings and also provide mappings, which are not present in BioPortal, the score decreases. In this case, we get the best results with the random forest machine learning approach, with a F1-measure of about 67%.



(a) TP of linear regression model using REST API

(b) TP of linear regression model using map file

(c) TP of random forest model using REST API

(d) TP of random forest model using map file

(e) TP of naive model using REST API

(f) TP of naive model using map file

Fig. 5: True Positives (TP) of different approaches prediction the quality of a match

Considering these results, we can say that our approach is able to assess the quality of mappings between two ontologies, especially if these matchings are based on lexical similarity. However, we have to admit that there is still a lot room for improvement. Therefore, we want to extend our framework in the future to also support metrics based on semantics to get better results. Furthermore, we aim to provide a framework which does not only consider two ontologies, but multiple. However, it is still hard to evaluate the quality of mappings between ontologies, because there are very few reference data sets, which provide a standard for the quality of matches. Therefore, it is also a goal in the future to provide such a reference data set.

6.1 Personal conclusions

Personal conclusion Tobias The main part I contributed to the project, was the handling of the information we retrieved from the ontologies ,i.e., how to use the REST API and later on, when we decided to use ontology files, to adapt our code to work with the *Apache Jena* library. Furthermore, I did research on the Children Metric and implemented it. Finally, I also evaluated the data we retrieved as output from our program. During the project, I have learned a lot about the structuring of data in an ontology, how it is represented in a file, and how we can access the information in it. Furthermore, one thing I learned is that handling such big amounts of data is really difficult, especially if data is not complete and only present in different formats.

Personal conclusion Lina Primarily, I contributed to applying lexical similarity metrics to the different properties. Also, I focused on the data handling when required, specially when retrieving properties from the ontology classes. In the initial phase of the project, I contributed with the problem definition and proposed solution, alongside with my colleagues. I also contributed to the literature research, specially with respect to the metrics. Finally, I contributed with the documentation during the meetings, as well as the writing of this report. I learnt a lot about the practical aspect of knowledge graphs (KG) and how to technically interact with them. Also, I learnt about information retrieval in the context of ontologies and KG, and the evaluation of results.

Personal conclusion Jonas My main contributions to the code base were focused on the loop detection, the application of the similarity metrics to subgraphs, the modelling of the data and development of helper functions whose need arose during development. Additionally, I spent time on literature research, mainly focusing on mapping evaluation and loop detection, and on the documentation of our progress. I learned a lot about knowledge graphs and ontologies in particular, and also realized that the application of the theory to practice is not always as easy as expected. For instance, there is nearly never an exact truth, but only different approximations, which makes the evaluation of your results much more difficult.

References

1. Noy, N., Mcguinness, D.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University (2001)
2. Kalfoglou, Y., Schorlemmer, M.: Ontology Mapping: The State of the Art. Knowl. Eng. Rev.**18**(1), 1–31 (2003)
3. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W541-5. Epub 2011 Jun 14.
4. BioPorltal - NCBO Wiki, `https://www.bioontology.org/wiki/BioPortal_Help`. Last accesed 04 Aug 2019
5. Stanford Knowledge Systems - AI Laboratory - What is an Ontology?, `http://www-ksl.stanford.edu/kst/what-is-an-ontology.html`. Last accessed 27 Jul 2019
6. Modulo 5.2 Ontology Matching - Knowledge and Data VU Amsterdam, `https://www.youtube.com/watch?v=gnq9IOOTjRo&t=920s`. Last accessed 27 Jul 2019
7. Hooi Y.K., Hassan M.F., Shariff A.M.: A Survey on Ontology Mapping Techniques. In: Jeong H., S. Obaidat M., Yen N., Park J. (eds.) Advances in Computer Science and its Applications 2014, Lecture Notes in Electrical Engineering, vol. 279. Springer, Berlin, Heidelberg (2014). https://doi.org/10.1007/978-3-642-41674-3_118
8. Euzenat, J., Shvaiko, P.: Ontology Matching. 2nd edn. Springer Publishing Company, Incorporated (2013)
9. Heflin, J., Song, D.: Ontology Instance Linking: Towards Interlinked Knowledge Graphs. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 4163–4169. AAAI Press, Phoenix, Arizona (2016)
10. Shvaiko, P., Euzenat, J.: Ten Challenges for Ontology Matching. In: Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. Part II on On the Move to Meaningful Internet Systems, pp. 1164–1182. Springer-Verlag, Monterrey, Mexico (2008)
11. Saveta, T., Daskalaki, E., Flouris, G., Fundulaki, I., Herschel, M., Ngonga Ngomo, A.: Pushing the Limits of Instance Matching Systems: A Semantics-Aware Benchmark for Linked Data. In: Proceedings of the 24th International Conference on World Wide Web, pp. 105–106. ACM, Florence, Italy (2015)
12. Euzenat, J.: Semantic Precision and Recall for Ontology Alignment Evaluation. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 348–353. Morgan Kaufmann Publishers Inc., Hyderabad, India (2007)
13. Jaccard Coefficient - an overview, `https://www.sciencedirect.com/topics/computer-science/jaccard-coefficient`. Last accessed 04 Aug 2019
14. Likavec, S., Lombardi, I., Cena, F.: How to improve Jaccard's feature-based similarity measure. In: Semantic Web Journal (2017)
15. Oliveira, D., Butt, A., Haller, A., Rebholz-Schuhmann, D., Sahay, R.: Where to search top-K biomedical ontologies?. In: Briefings in Bioinformatics. (2018)
16. Pathak, J., Chute, C.: Debugging mappings between biomedical ontologies: Preliminary results from the NCBO BioPortal mapping repository. In: ICBO, pp. 95. Citeseer, (2009)
17. Jean-Mary YR, Shironoshita EP, Kabuka MR: Ontology Matching with Semantic Verification. In: Web Semant. 2009;7(3):235–251. (2009)
18. Y. A. Effendi, R. Sarno: SWRL rules for identifying short loops in business process ontology model. In: 2017 11th International Conference on Information & Communication Technology and System (ICTS), Surabaya, 2017, pp. 209-214. (2017)

19. Mehdi Teymourlouie, Ahmad Zaeri, Mohammadali Nematbakhsh, Matthias Thimm, Steffen Staab. Detecting hidden errors in an ontology using contextual knowledge. In: Expert Systems with Applications, Volume 95, 2018, Pages 312-323. (2018)
20. Achichi, M., Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., Flouris, G., Fundulaki, I., Harrow, I., Ivanova, V., Jiménez-Ruiz, E., Kolthoff, K., Kuss, E., Lambrix, P., Leopold, H., Li, H., Meilicke, C., Mohammadi, M., Montanelli, S., Pesquita, C., Saveta, T., Shvaiko, P., Splendiani, A., Stuckenschmidt, H., Thiéblin, É., Todorov, K., Trojahn dos Santos, C., Zamazal, O.: Results of the Ontology Alignment Evaluation Initiative 2017. In: OM: Ontology Matching (CEUR Workshop Proceedings), pp. 61-113., Wien, Austria (2017)
21. 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. In: J. Biomedical Semantics, pp. 55.BioMed Central, (2017)
22. Fleischhacker, D., Stuckenschmidt, H., A.: A Practical Implementation of Semantic Precision and Recall. In: 2010 International Conference on Complex, Intelligent and Software Intensive Systems, pp. 986-991.(2010)
23. Amith, M., He, Z., Bian, J., Lossio-Ventura, J. A., Tao, C.: Assessing the practice of biomedical ontology evaluation: Gaps and opportunitiesy. In: Journal of biomedical informatics, pp. 1–13. (2018)
24. Todorov, K., Geibel, P., Kühnberger, K.: Mining Concept Similarities for Heterogeneous Ontologies. In: Proceedings of the 10th Industrial Conference on Advances in Data Mining: Applications and Theoretical Aspects, pp. 86–100. Springer-Verlag, Berlin, Germany. (2010)