Too Big to Match: a Strategy Around Matching Tasks for Large Taxonomies

Alsayed Algergawy[1,*], Naouel Karam[2], Amir Laadhar[3] and Frank Michel[4]

[1]Institute for Computer Science, University of Jena, Germany
[2]Fraunhofer FOKUS, InfAI e.V., Berlin, Germany
[3]University of Stuttgart, Germany
[4]University Côte d'Azur, CNRS, Inria, I3S, Nice, France

Abstract

Following the introduction of a new matching task at the biodiversity and ecology track of the Ontology Alignment Evaluation Initiative (OAEI), to align between two large taxonomies, we acknowledged the fact that large ontologies or taxonomies still cannot be efficiently tackled by state-of-the-art ontology matching systems. In this paper, we take advantage of structural specificities of taxonomies to devise a strategy for deviding large scale taxonomy matching tasks into smaller, more manageable subtasks. Our modularization approach is based on a locality-based module extraction technique. We conducted a first assessment of the coverage of the obtained modules as well as a preliminary evaluation using a set of tools from OAEI.

Keywords

Ontology matching, Large taxonomies, NCBITAXON, TAXREF-LD

1. Introduction

At the 2021 edition of the Ontology Alignment Evaluation Initiative (OAEI), we introduced a new matching task to align between the NCBI Organismal Classification [1] (NCBITaxonomy for short) and TAXREF-LD [2]. Both are large biologic taxonomies that respectively contain 1,983,907 and 285,863 classes. No matching system succeeded in the matching of NCBITaxonomy and TAXREF-LD, in the given time frame and with constrained computing resources. Indeed, the upmost search space that need to be considered by matching systems is the cartesian product of entities from the two input taxonomies, which for the task at hand represents almost 529 billion candidate correspondences. The same issue has been addressed by the OAEI Large BioMed Track (largebio) for many years, the size of the biggest considered ontology corresponding to the range of TAXREF-LD with around 300,000 classes. The largebio track organizers proposed and applied a strategy to divide the ontologies into smaller fragments [3], the track consists of manageable subtasks with different fragments sizes. Furthermore, several approaches have

OM2022: the International Workshop on Ontology Matching, October 23, 2022, Hangzhou, China

*Corresponding author.
alsayed.algergawy@uni-jena.de (A. Algergawy); naouel.karam@fokus.fraunhofer.de (N. Karam); amirl@cs.aau.dk (A. Laadhar); fmichel@i3s.unice.fr (F. Michel)
0000-0002-8550-4720 (A. Algergawy); 0000-0002-0877-7063 (N. Karam); 0000-0001-7116-9338 (A. Laadhar); 0000-0002-9421-8566 (F. Michel)

been proposed in the literature to support matching large ontologies [4, 5, 6], e.g. reduction of the search space and parallel matching are two common strategies. A state of the art and a comparison of the approaches based on the employed technique has been presented in [7].

In this paper, we propose a method to split the large matching task into a set of smaller, more manageable subtasks through the use of modularization. Each subtask will cover one self-contained module from each of the two taxonomies. Unlike state of the art approaches that operate on ontologies and hence need to deal with issues around preserving the coverage of the relevant ontology alignments, as well as an accurate and self-contained division, our approach is tailored for taxonomies and their specificities. Indeed, structural characteristics of biological taxonomies allow us to divide based on a specific hierarchy level, the so-called taxonomic rank and thus obtain independent and logically relevant modules. We performed a preliminary evaluation of the new matching tasks based on the obtained taxonomy modules and their corresponding subset of the reference alignments using a set of matching systems from OAEI.

2. Characteristics of Biological Taxonomies

In this section, we describe some characteristics of biological taxonomies and their particular structure that serves as basis for our modularization approach.

Large size. Biologic taxonomies are used to name, define and classify groups of biological organisms based on common characteristics. They result from a huge amount of efforts that taxonomists have dedicated in studying living organisms and publishing their knowledge for over two centuries. Those are in consequence inherently large. For instance, at the time of writing this paper, the NCBI taxonomy contains over 2 million species. TAXREF-LD, although restricted to the french territories, still counts at this day over 200,000 species. In addition to the constant growth in size, the progress of species studies leads to frequent changes like the redefinition, reclassification and merging of taxon concepts, making taxonomies highly dynamic [8].

Structure and naming conventions. Biological taxonomies consist of a hierarchy where each level is assigned a taxonomic rank, such as species, genus, family, etc. up to kingdom. In the hierarchy, the parent of a taxon is the taxon with higher taxonomic rank (e.g. ”Delphinus delphis” is of rank species whereas its parent class ”Delphinus” is of rank genus). Different taxonomies adopt certain perspective, or are meant for certain purpose, thus covering complementary and possibly overlapping regions, epochs or domains. For instance, NCBITaxonomy covers the organisms referenced in the NCBI nucleotide and protein sequences database. Biologic taxonomies consist of two coexisting yet distinct realities, the taxonomy (the description and characterization of biological entities called taxa), and the nomenclature (the rules specifying how to define scientific names and assign them to taxa).

High dynamicity. The circumscription of a taxon, i.e. the set of individuals that it actually consists of, is provided by the set of names that may be used to refer to it: the reference name (called *accepted name* in zoology or *valid name* in botany), and possibly multiple synonyms. These multiple names stem from the fact that scientific consensus constantly evolves, in light e.g. of new evidence, that leads to recombination of taxa (merging, splitting, moving from one parent to another etc.). The way these recombinations must be handled is specified by nomenclatural rules that are compiled in several *Codes of nomenclature*, for animals, for plants and fungi, and for bacteria.

3. Alignment Task at Hand

The NCBI Taxonomy is the standard nomenclature and classification repository for the source organisms in the sequence databases of the International Nucleotide Sequence Database Collaboration (INSDC). The NCBITaxonomy ontology is an automatic translation of the NCBI taxonomy database into OWL. The translation treats each taxon as a class whose instances would be individual organisms. The NCBI Taxonomy is updated daily but the releases of the OWL counterpart are triggered manually by an OBO administrator bi-annually. The OAEI biodiv track uses the release 2021-02-15, containing 1,983,907 classes.

TAXREF is the French taxonomic register for fauna, flora and fungus, maintained and curated by the National Museum of Natural History of Paris. TAXREF is available in multiple formats, in particular as a knowledge graph based on the linked data principles, called TAXREF-LD [2]. To account for the distinction between taxonomy and nomenclature , TAXREF-LD holds two distinct levels of modeling. At the taxonomic level, each biological taxon is modeled as an OWL class whose members are the biological individuals in that taxon. At the nomenclatural level, scientific names are represented as the concepts of a SKOS thesaurus. The OAEI biodiv task currently relies on TAXREF-LD version 13 that registers 266,846 taxa and 657,609 scientific names. Since the tools cannot deal with OWL and SKOS at the same time, we turned the SKOS part into simple name labels attached to the OWL classes.

TAXREF-LD comes with alignments to NCBITaxonomy, computed using the SILK framework [9], that the authors extended with a plugin that implements rules for the alignment of scientific names. These rules are designed to work around common mistakes that are being made when spelling scientific names. This typically pertains to the use of parentheses and abbreviations, accentuated characters, or the transcription of letters using the Latin ligature (e.g. "Æ" may be spelled "AE"). Furthermore, only taxa of species or infra-specific ranks were considered to compute the reference alignments. A reason is that this is where the vast majority of taxa are. A more pragmatic reason is that the names of species or infra-specific ranks consist of at least two terms, the genus followed by an epithet, which is quite discriminating. Conversely, names in higher ranks are usually single-worded and some of them may be very similar, sometimes varying by only one letter (e.g. sub-family *Tenrecinae* belongs to family *Tenrecidae*), such that lexical alignment methods tend to produce lots of false-positives.

4. Modularization of the Taxonomies

In general, there are two ways to split an ontology into a set of partitions: *module extraction* and *module partitioning* [10], where module extraction aims to extract from the given ontology a small fragment that captures the intended meaning of input terms [11, 12], while ontology partitioning splits the given ontology into a set of modules. In this work, we make use of a module extraction technique.

Locality-based module extraction. The locality-based module extraction is the process that extracts a meaningful subset of an ontology given a number of terms (signature). The extracted module guarantees to completely capture the meaning of the given set of terms. In the context of the alignment task at hand, we first prepared the set of terms that will be used as input for the module extraction process. The current implementation of the locality-based approach supports extracting three types of syntactic-locality-based modules: bottom module, top module and the star module [10]. In this work, we applied the *locality-bottom and -top module extraction strategies.* Thus covering the relevant hierarchy information needed by matching systems.

Input terms preparation. The specific structure of the taxonomies hierarchy based on taxonomic rank levels described in Section 2 constitute the perfect basis to divide the taxonomy into meaningful and independent modules. We chose to start at the highest taxonomic rank, namely the kingdom. We made use of the TAXREF-LD API to extract the kingdom of all entities appearing in the original set of reference alignments, then we grouped the alignments by kingdom. We obtained 6 groups corresponding to the kingdoms: Animalia, Bacteria, Chromista, Fungi, Plantae and Protozoa. We then used the obtained set of terms together with their original taxonomy as input for the locality-based module extraction tool.

After applying the locality-based module extraction on both taxonomies given the input set of terms for each category, we get six modules for each taxonomy. The number of concepts within each module is presented in Table 1. The matching task has been split into six well balanced matching subtasks. This is due to the nature of the extraction process, since we used the same set of input terms and the taxonomies share a similar rank-based structure.

Table 1
Sizes of obtained modules for each taxonomy

Matching task	T1 Animalia	T2 Bacteria	T3 Chromista	T4 Fungi	T5 Plantae	T6 Protozoa
NCIBITAXON	74729	326	2344	13149	27013	538
TAXREFLD	73528	312	2290	12732	26302	501

In prevision of this year's OAEI edition, we conducted a preliminary evaluation of the obtained modules using a set of OAEI matching tools. We ran three different matching systems, namely AML [13], LogMap [14] and ATBox [15]. Our goal with this first evaluation was to ensure matching systems will be able to deal with all subtasks. All systems completed the tasks successfully. We will perform a full evaluation of the participating systems of this year's edition based on the OAEI schedule.

In Table 2, we show the number of mappings computed by each system for each subtask. All systems computed nearly the same size of mappings, these are however much bigger than our reference alignments. This is probably due to the fact that we considered only species or infra-specific ranks (c.f. Section 3). For the final evaluation, we will ignore the set of mappings not covered by the reference alignment and perform a manual assessment of the mappings produced by 2 systems or more, to be potentially added to our reference alignment.

Table 2
Number of computed mappings by each system on the matching subtasks

Task (size of reference)	T1 (48220)	T2 (175)	T3 (1405)	T4 (10162)	T5 (19914)	T6 (357)
AML	71269	303	2219	12937	26671	496
LogMap	72838	302	2219	12937	26862	496
ATBox	71383	295	2192	12623	25862	478

5. Conclusion

We have presented an approach to divide a taxonomy matching task into subtasks based on taxonomic ranks and a locality-based module extraction. The obtained modules are consequently logically coherent and independent from each other. The bottom-up module extraction strategy guarantees the inclusion of relevant information required by matching systems while ensuring the coverage of the initial reference alignment. We tested a set of systems on the sub-tasks and we will be performing a full evaluation at the 2022 OAEI edition.

Acknowledgements

This work has been partially funded by the German Research Foundation (DFG) as part of the CRC 1076 Aquadiva, NFDI4Biodiversity (442032008) and NFDI-MatWerk (460247524) projects.

References

[1] S. Federhen, The NCBI Taxonomy database, Nucleic Acids Research 40 (2012).

[2] F. Michel, O. Gargominy, S. Tercerie, C. Faron-Zucker, A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF, in: Proceedings of the ISWC2017 workshop on Semantics for Biodiversity (S4BioDiv), volume 1933, Vienna, Austria, 2017.

[3] E. Jiménez-Ruiz, A. Agibetov, J. Chen, M. Samwald, V. Cross, Dividing the ontology alignment task with semantic embeddings and logic-based modules, in: ECAI 2020, Santiago de Compostela, Spain, volume 325, 2020.

[4] E. Rahm, Towards large-scale schema and ontology matching, in: Schema matching and mapping, 2011.

[5] F. Hamdi, B. Safar, C. Reynaud, H. Zargayouna, Alignment-Based Partitioning of Large-Scale Ontologies, 2010.

[6] W. Hu, Y. Qu, G. Cheng, Matching large ontologies: A divide-and-conquer approach, Data & Knowledge Engineering 67 (2008).

[7] P. Ochieng, S. Kyanda, Large-scale ontology matching: State-of-the-art analysis, ACM Computing Surveys (CSUR) 51 (2018).

[8] A. Kohlbecker, N. Karam, A. Paschke, A. Güntsch, Preserving taxonomic change and subsequent taxon relationships over time, in: Proceedings of the Joint Ontology Workshops 2021 Episode VII: The Bolzano Summer of Knowledge, volume 2969, 2021.

[9] J. Volz, C. Bizer, M. Gaedke, G. Kobilarov, Silk - A Link Discovery Framework for the Web of Data., in: 2nd Workshop about Linked Data on the Web, Madrid, Spain, 2009.

[10] A. Algergawy, S. Babalou, F. Klan, B. König-Ries, Ontology modularization with oapt, Journal on Data Semantics (2020).

[11] B. C. Grau, I. Horrocks, Y. Kazakov, U. Sattler, Just the right amount: extracting modules from ontologies, in: Proceedings of the 16th international conference on World Wide Web, 2007.

[12] A. A. Romero, M. Kaminski, B. C. Grau, I. Horrocks, Module extraction in expressive ontology languages via datalog reasoning, J. Artif. Intell. Res. (2016).

[13] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, F. M. Couto, The agreementmakerlight ontology matching system, in: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", 2013.

[14] E. Jiménez-Ruiz, B. Cuenca Grau, Logmap: Logic-based and scalable ontology matching, in: International Semantic Web Conference, 2011.

[15] S. Hertling, H. Paulheim, Atbox results for oaei 2020, in: CEUR Workshop Proceedings, volume 2788, 2020.