Comparison of ontology mapping techniques to map plant trait ontologies

Marie-Angélique Laporte, Léo Valette, Elizabeth Arnaud
Bioversity International
Montpellier, France
m.a.laporte@cgiar.org

Laurel Cooper, Austin Meier, Pankaj Jaiswal
Department of Botany and Plant Pathology
Oregon State University
Corvallis, USA

Christopher J. Mungall
Berkeley Bioinformatics Open-Source Projects
Lawrence Berkeley National Laboratory
Berkeley, USA

***Abstract*—Crop specific ontologies for phenotype annotations in breeding have proliferated over the last 10 years. Across-crop data interoperability involves linking those ontologies together. For this purpose, the Planteome project is mapping the Crop Ontology traits (www.cropontology.org) to the reference ontology for plant traits, Trait Ontology (TO). Manual mapping is time-consuming and not sustainable in the long-run as ontologies keep on evolving and multiplicating. We are thus working on developing reliable automated mapping techniques to assist curators in performing semantic integration. Our study shows the benefit of the ontology matching technique based on formal definitions and shared ontology design patterns, compared to standard automatic ontology matching algorithm, such as AML (AgreementMakerLight).**

Keywords—ontology mapping; ontology design patterns; reference ontologies

I. INTRODUCTION

The development of improved crop varieties relies on both traditional breeding methods and next-generation methods such as high-throughput sequencing, molecular breeding and automated scoring of traits. In that context, a number of ontologies have been developed to face the data interoperability issues. They fulfill the needs of specific communities, but are species or clade-specific ontologies [1] and therefore block data harmonization across disciplines and communities.

The crop breeding community, in particular widely uses the Crop Ontology (CO; www.cropontology.org), which is composed of species-specific ontologies for fieldbook edition and data annotation [1]. Because these ontologies grow in size and number, it is essential to develop efficient and reliable automated concept mapping techniques to be able to apply semantics channels for data integration and discovery.

The Planteome project (www.planteome.org) aims to support comparative plant biology, and provides integrated access to annotated datasets generated by inter and intra-specific comparative analysis of transcriptomes, proteomics, phenomics and genome annotation. To address this objective, Planteome is currently developing, and promoting the use of a set of reference ontologies for plants, proposing species-neutral concepts, as well as common data annotation standards. Harmonization between the species-specific ontologies and the Planteome reference ontologies is currently done by mapping Crop Ontology to the Plant Trait Ontology (TO) [2], which is the reference species-neutral ontology for plant traits aiming at integrating the many crop-specific trait ontologies.

The purpose our study is to generate mappings in an efficient way in order to ease the work of the ontology curators in creating manual mappings. In this objective, we have compared two automatic ontology mapping techniques. The first technique is widely used to align ontologies and consists in applying a standard automatic matching algorithm. Indeed, AML (AgreementMakerLight) performs mappings based on both the string similarities of the ontology terms and the ontology structure. Considering the number of ontologies to be mapped and the inherent nature of ontologies to evolve over time, it can be hard to maintain automatically the mappings created using such a technique. Therefore, the second technique relies on formal definitions and shared ontology design patterns. The ontology design patterns are created using Ontology Web Language (OWL) axioms based on Entity-Quality (EQ) statements, leading to a post-composition of terms, similar to what has been proposed by the Ontology of Biological Attributes (OBA) [3]. The Entity (E) and Quality (Q) are sourced from the reference ontologies promoted by

Planteome. The Q comes from the Phenotype and Trait Ontology (PATO) whereas the E comes from Plant Ontology (PO) when it is related to plant structures, Gene Ontology (GO) for subcellular components, Chemical Entities of Biological Interest (ChEBI) for chemical entities or Environment Ontology (EO) for the environment conditions. Automated reasoning engines are then used to generate the mappings between the species-specific ontologies and the reference ontologies, while guarantying the validity of the unified merged ontology (i.e. TO plus the species-specific CO). As a result, TO is being enriched with well defined crop-specific terms of Crop Ontology and Planteome can integrate additional data annotated in a unified way by the breeding and the genetic communities.

II. RESULTS AND DISCUSSION

The AML algorithm and the design patterns approach have been applied to four crop Trait Dictionaries of the Crop Ontology so far: cereals rice and wheat, legume lentil and root tuber crop cassava. Those ontologies are very different in terms of plant anatomy and morphology, but also in terms of count and complexity of phenotypic traits. Table 2 summarizes the results of the mappings process on trait terms. Mapping using formal definitions resulted in two-fold increase successful mappings. On average, AML was able to propose mappings for ~40% of the CO classes in each ontology compared to ~75% mapped terms using the formal definition approach. This can be explained by the fact that crop specific ontologies use very specific terminologies, especially for the Entity part of the EQ statement. Although the specific plant entities are defined in the Plant Ontology (PO) as synonyms of species neutral entities, all the synonyms were not added to TO and CO when the terms were pre-composed. The AML algorithm was thus not able to use this information, whereas the PO synonyms have been used in order to build the formal definitions of the CO terms. Furthermore, because the class hierarchy is quite simple in the different CO, AML was not able to use the ontology structures to improve the mapping results: only equivalent terms were found using AML.

Disease resistance traits are important for breeders. A disease results from the combination of a host species, a pathogen and an environment, the disease resistance traits are crop-specific. Biotic stress traits include disease-related traits and can cover as much as 20% of the individual CO. Those traits cannot have an exact correspondence in TO. Thus AML was not able to find mappings for those terms. Based on the formal definitions, a reasoner linked those terms to be subclasses of one the TO stress trait.

Finally, all the classes in TO haven't been formally defined. Indeed, design patterns are hard to develop for very complex traits such as yield-related traits. This is why the all the CO classes cannot be mapped to TO classes using the design pattern technique. Manual mapping is still needed in order to map those traits. The mapping coverage will be improved in the future. The mapped ontologies are available on www.planteome.org as well as on Planteom's GitHub repository (https://github.com/Planteome).

TABLE I. MAPPING RESULTS

	Rice	**Wheat**	**Lentil**	**Cassava**
# trait classes	157	238	66	175
AML	84 (54%)	73 (30%)	28 (42%)	59 (34%)
Design Patterns	121 (77%)	199 (84%)	47 (71%)	118 (67%)

III. CONCLUSION

In an era of ontology proliferation, it is of vital importance to have reference ontologies and powerful tools that reduce the effort of ontology alignment. Standard mapping techniques do not fit the need of ontology evolution over time as their results are difficult to maintain automatically. Developing the mapping process based on ontology design patterns and logical axioms ensures validity confidence accuracy of the resulting ontology mappings. Scientists from the breeding community can continue to use the standards preferred by them to annotate/record their data, reducing the effort they need to provide. Planteome, through the TO, provides unified access to the breeding and the genetic data, opening up the possibility to perform large scale analysis such as comparative genomics by promoting a species neutral approach.

ACKNOWLEDGMENT

This work was supported by IOS:1340112 from the NSF. Additionally, CJM acknowledges the support of the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

REFERENCES

[1] Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G.F., Hancock, D., Morrison, N., Bruskiewich, R. and McLaren, G., 2010. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB plants*, *2010*, p.plq008.

[2] Arnaud, E., Cooper, L., Shrestha, R., Menda, N., Nelson, R.T., Matteis, L., Skofic, M., Bastow, R., Jaiswal, P., Mueller, L.A. and McLaren, G., 2012, October. Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes. In *KEOD* (pp. 220-225).

[3] https://github.com/obophenotype/bio-attribute-ontology, DOI:10.5281/zenodo.47337