Analysis of Term Reuse, Term Overlap and Extracted Mappings across AgroPortal Semantic Resources

Amir Laadhar, Elcio Abrahão, Clement Jonquet

Analysis of Term Reuse, Term Overlap and Extracted Mappings across AgroPortal Semantic Resources

Amir Laadhar, Elcio Abrahão[0000−0001−7983−2253], and Clement Jonquet[0000−0002−2404−1582]

Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), University of Montpellier & CNRS, France
firstname.lastname@lirmm.fr

Abstract. Ontologies in agronomy facilitate data integration, information exchange, search and query of agronomic data, and other knowledge-intensive tasks. We have developed AgroPortal, an open community-based repository of agronomy and related domains semantic resources. From a corpus of ontologies, terminologies, and thesauri taken from AgroPortal, we have generated, extracted and analyzed more than 400,000 mappings between concepts based on: (i) reuse of the same URI between concepts in different resources –term reuse; (ii) lexical similarity of concept names and synonyms –term overlap; and (iii) declared mappings properties between concepts –extracted mappings. We developed an interactive visualization of each mapping construct separately and combined which helps users identify most prominent ontologies, relevant thematic clusters, areas of a domain that are not well covered, and pertinent ontologies as background knowledge. By comparing the size of the semantic resources to the number of their mappings, we found that most of them have under 5% of their terms mapped. Our results show the need of an ontology alignment framework in AgroPortal where mappings between semantic resources will be assembled, compared, analysed and automatically updated when semantic resources evolve.

Keywords: Ontology alignment · Term reuse · Term overlap · Extracted mappings · Mapping analysis · Visualization

1 Introduction

By reusing the NCBO BioPortal technology [16], we have designed AgroPortal (`http://agroportal.lirmm.fr`), an ontology repository for agronomy, plant and food sciences and originally biodiversity-ecology [9]. As of August 2020, AgroPortal includes 126 ontologies, terminologies and thesauri encoded in different formats like RDFS, OWL, OBO, UMLS-RRF and SKOS. AgroPortal stores reference resources such as the Plant Ontology or Agronomy Ontology or AGROVOC. The need for interconnecting these resources i.e., ontology alignment [3], has been explicitly expressed by almost all of our partners and collaborators to achieve interoperability among their semantic resources. But the need

goes beyond the sole ability to automatically generate alignment between ontologies, it includes being able to store, compare, evaluate the mappings. Therefore, AgroPortal offers a mapping repository to store mappings between its semantic resources. To build this mapping repository, we currently consider three mappings constructs:[1] term reuse, term overlap and extracted mappings. By *term reuse*, we mean the situation in which a term of an ontology is explicitly reused inside another ontology using the same URI.[2] Term reuse is a good practice in ontology/terminology development as it facilitates semantic interoperability and reduce ontology engineering efforts [1]. However, for many reasons, it is not a common practice when semantic resources are not developed under the same umbrella or by the same group or simply when an ontology developer likes to add statements to an object which he/she does not want to conflict with statements in other ontologies. By *term overlap*, we mean the situation in which two classes/concepts use the same labels or synonyms in different semantic resources. Lexical matches are clearly known not to be fully reliable as semantic mappings simply because of the polysemic aspects of labels. However, they are also very well perceived as a useful and quick way of finding relevant similar concepts/ontologies [5]. By *extracted mapping*, we mean being able to extract and load in the repository mappings explicitly declared inside the ontology source files (typically using `owl:sameAs` or SKOS mapping properties) to reify them into first-class objects with provenance information. Contrary to expectations, the process of extracting mappings is not trivial considering the heterogeneity of means to encode mappings and the predominant use of ambiguous constructs like OBO XRefs for instance.

AgroPortal's mapping repository is valuable to our community, since it allows ontology developers and users to identify similar terms across ontologies and it facilitates data integration in systems relying on different semantic resources. Mappings help the identification of prominent ontologies that can serve as a common denominator or hub for data interoperability. If AgroPortal easily detects term reuse between ontologies, the identification of correct term overlap is harder because of polysemic labels and can bring to incoherences [4, 17]. However, these "overlaps" are very useful as they can be used by developers to identify similar or equivalent terms to manually enrich their ontologies by declaring formal and rigorous mappings. Today, term reuse and term overlap are automatically detected by AgroPortal when an ontology is uploaded. However, we are currently working to automate mapping extraction from files during the ontology parsing routine.

In this article, we present an analysis of the mapping repository on a corpus of 109 ontologies built from AgroPortal's content in March 2020. Such analysis of the mappings between semantic resources is important as it tells us about the

[1] We prefer here the term "construct" to "type" which is used in our work with another meaning: to quality the mapping (exact match, close match, same as, etc.).

[2] The most frequent case of reuse concern classes/concepts, however any object identified by an URI can be reused from one semantic resource to the other (e.g., `owl:Class`, `owl:Individual`, `rdfs:Property`, `skos:Concept`).

landscape of semantic resources, the structure of the ontology repository, and the ways mappings can help in the process of ontology design and evaluation. The contributions of this work are the following:

- A dataset of multiple mappings constructs between semantic resources in agronomy and related domains which is in large part curated;
- An openly available tool called Ontology Mapping Harvester Tool (OMHT), which automatically extracts mappings declared inside ontology source files and represent them into classic mapping formats;
- An interactive visualization of the mapping dataset to display mapping constructs individually and combined;
- A descriptive analysis for each mapping construct.

The rest of the paper is organized as follows: the next section presents related work. In Section 3, we introduce the methodology used for each of the three mapping constructs for the generation the mappings dataset. In Section 4, we describe the analysis and introduce the visualization. Finally, in section 5, we discuss our results and conclude in Section 6.

2 Related Work

Ontology alignment is a key aspect for ontologies: it makes them more interoperable and interconnect the ones on overlapping domain of interests [3]. There is lack of reference mapping repositories that would serve mappings as FAIR data [21]. Such mappings repositories should support representation, extraction, harvesting, generation, validation, merging, evaluation, visualization, storage and retrieval of mappings between the ontologies they host and other ones [8].

Mappings are handled differently within ontology repositories: Ontohub[19], for instance, allows browsing, searching, and aligning ontologies. To the best of our knowledge, only the repositories in the OntoPortal family[3] offers an integrated mapping repository, where different kind of mappings are stored with provenance and are accessible (read/write) through the user interface or via API calls. BioPortal [16] automatically detects term reuse and generates term overlap mappings (with a method called LOOM [5]), however, the technology does not embed any state-of-the-art automatic ontology matching systems and does not extract mappings declared inside the ontologies to reify them inside the mapping repository. Other initiatives, such as the UMLS Metathesaurus includes a specific table to store mappings between the medical terminologies (MRMAP). The European Bioinformatics Institute develops OxO (Ontology Xref Service) to visualize cross-references mappings (i.e., declared with the `oboInOwl:hasDbXref` property) extracted from ontologies inside the Ontology Lookup Service [10]. To disambiguate the prefix of XRefs targets and identify data sources, OxO uses Identifiers.org, the OBO Library, and Prefixcommons.org. Whereas in our work,

[3] The NCBO BioPortal technology can be reused and customized for deploying other ontology repositories e.g., AgroPortal or EcoPortal. Since 2019, the generic technology is branded as OntoPortal (`https://ontoportal.org`)

we semi-automatically curate the declared cross-references to keep only explicit valid mappings between ontology terms [13].

Mapping repositories are useful for several applications such as modules extraction, ontology partitioning, ontology alignment using background knowledge resources and mappings visualization. For instances: Amina et al. [2] proposed a background knowledge-based ontology matching system using as background knowledge a graph, build-out of external mappings, to interconnect the source and target ontologies and identify mapping candidates. Ghazvinian et al. [6] used mappings to extract modules from large ontologies. The YAM++ ontology matching system [15] defined a machine learning classifier trained on a set of reference external mappings. Kamdar et al. [12] proposed a visualization for mappings extracted from BioPortal but it was not maintained in sync with the ontology repository after publication.

Similar mappings analysis work to the one presented in this article are: In 2009, Ghazvinian et al. [7] analyzed more than four million term overlap mappings between 200 ontologies or terminologies in BioPortal (including 67 terminologies from the UMLS Metathesaurus). The mappings were generated with a simple lexical matching method to identify classes with same labels preferred terms and synonyms, over normalised strings. Although their approach was technically simple, they have demonstrated the value of the mappings extracted [5]. They performed term overlap analysis to learn more about the characteristics of the ontologies and the relationships between them e.g., identify hubs and clusters over the ontologies. They used network analysis methods to answer practical questions and to reason about the distribution of mappings among the ontologies. In 2012, Poveda et al. [18] analyzed the landscape of reuses in the 196 semantic resources included in the Linked Open Vocabularies (LOV) registry[20]. In 2015, Kamdar et al.[11] investigated term reuse and overlap in 377 biomedical ontologies from BioPortal. However, in this study, XRef mappings were mixed with other term reuses whereas in our work we distinguish them and consider them as extracted mappings. The authors highlighted the need for a sophisticated term recommendation mechanisms that support consistent term reuse. Later, the authors extend their study to 509 ontologies in BioPortal and reported a term reuse of 9% and a term overlap of 22.23% [12].

3 Methodology: Mapping Dataset Creation

We used for this study a set of 109 distinct agri-food and biodiversity-ecology semantic resources, hosted in AgroPortal in March 2020. These semantic resources include 9 ontologies in the OBO format (e.g., SOY, GR-TAX), 88 OWL ontologies (e.g., FOODON, ATOL), 10 SKOS thesauri (e.g., AGROVOC, NALT). 103 and 104 semantic resources (95%) show respectively term reuse and overlap; 28 semantic resources (25,68%) contain declared mappings in their source files.

In AgroPortal, term overlap and reuses are automatically detected but the system does not explicitly materialize these mappings with provenance and a

mapping relation.[4] However, when we build our corpus, we represent these mappings as any other mappings in the repository and assign them provenance information and relevant relations: `owl:sameAs` for term reuse and `skos:relatedMatch` for term overlap. Indeed, `skos:relatedMatch` is in SKOS the "weaker" mapping relation which is appropriate, we believe, for non-curated lexical mappings even if in some case `skos:exactMatch` or `skos:closeMatch` would be more appropriate. Extracted mappings already have a mapping property chosen by the ontology developer when he/she created the mapping. In our dataset, each mapping is also described with some metadata information using a BioPortal specific JSON mapping format e.g., creation date, creator/tool, comment. More detail about the creation of our dataset is provided in the next subsections.

We consider term overlap mappings as symmetric because when there is a match between two labels of two different ontologies, this match is independent of the source and target ontologies. Therefore, term overlap mappings are bidirectional. Even if the URI of the mapped entity is the same, we do not consider term reuse mappings symmetric since they explicitly state that an ontology reuses another one but not the other way around. Therefore, term reuse mappings are unidirectional. Similarly, extracted mappings can be considered symmetric or not depending on their semantics; but in this study, those mappings are being explicitly declared in one ontology source file and not necessarily in the other, we thus consider them unidirectional.

3.1 Term Reuse Mappings Harvesting

We define term reuse as the situation in which an URI from one ontology is explicitly reused inside another ontology. This situation occurs when the developers of semantic resources decide to rely on knowledge described in other resources. It increases the reusability between ontologies and reduces development time and the proliferation of equivalent terms. A developer can either decide to reuse specific terms one by one by simply identifying them with their URIs in a statement or by (re)declaring them locally using the original URI. Or he/she can import all the objects and statements of an ontology (or ontology module) into another one. The later is only possible with OWL ontologies using the construct `owl:imports`. Ontology developers typically use this construct to import ontology modules. Among 88 OWL ontologies, we found 15 of them that use `owl:imports`. For instance, the Food Ontology (FOODON) imports some modules from ENVO or ChEBI. The imported modules may themselves contain any kind of mappings like term overlap, term reuse and extracted mappings. Therefore, we include the set of imported mappings in our mappings dataset.

We obtain term reuse mappings from AgroPortal's REST API.[5] AgroPortal creates mappings between any two classes or concepts explicitly declared or imported from one ontology to another using the same URI. Those "same URI

[4] See `https://github.com/agroportal/documentation/wiki/Mappings` for details.

[5] E.g., the following call returns all the mappings between the Agronomy and Plant Ontology: `http://data.agroportal.lirmm.fr/mappings?ontologies=AGRO,PO`

mappings" are not materialized in the repository but generated on-the-fly. Several ontologies (especially in the OBO community) reuse multiple terms from one another, thus, from all these "same URI mappings", we keep only the ones corresponding to direct reuses from one ontology to the other. For example, AGRO's term "life of whole plant stage" originally defined in the Plant Ontology (PO:0025337) is same URI-mapped in AgroPortal to any other ontology using this term (e.g., PO, FLOPO, ENVO), however, we only retain as a term reuse the AGRO-PO, FLOPO-PO, ENVO-PO mappings. In addition, when an ontology reuses a term from another ontology, not in our corpus, we ignore this reuse. From our corpus of ontologies, we harvested a total of 16,958 term reuse mappings (over a total of more than 53,000 "same URI mappings").

3.2 Term Overlap Mappings Harvesting

We define term overlap as the situation in which two terms use the same labels or synonyms in different semantic resources. LOOM is an automatic ontology matching system [5] implemented in the OntoPortal technology –thus available in AgroPortal– to generate lexical matches between all the semantic resources independently of their original formats. To identify the correspondences, LOOM compares preferred names and synonyms of the terms in source and target ontologies and create a match, if and only if their labels are equal based on a modified-string comparison function. The tool first removes all delimiters from both strings (e.g., spaces, underscores, parentheses, etc.) and the accents. Then it uses an approximate matching technique to compare the strings, allowing for a mismatch of at most one character in strings with length greater than four and no mismatches for shorter strings.

We also obtain term overlap mappings from AgroPortal's REST API. Those "LOOM mappings" are also not materialized in the repository but generated on-the-fly. In AgroPortal, term overlaps are identified for any terms, being it reused from another ontology or not. Thus, from all these "LOOM mappings", we remove the ones corresponding to direct reuses from one ontology to the other. For example, AGRO's term "life of whole plant stage" originally defined in the Plant Ontology (PO:0025337) is Loom-mapped in AgroPortal to any other ontology using the same label (e.g., PO, FLOPO, ENVO), however, we only retain as a term overlap the AGRO-FLOPO, AGRO-ENVO and FLOPO-ENVO mappings. From our corpus of ontologies, we harvested a total of 246,348 term overlap mappings. Due to the large size of the harvested tern overlap mappings, we did not curate these mappings.

3.3 Extracted Mappings Harvesting

We mean by extracted mapping the ones explicitly declared inside the ontology source files and extracted to be to reified into a first-class objects with provenance information in a mapping repository or in our case included in our dataset. Extracted mappings are very valuable as they are usually manually created or curated by the ontology developers and because they are semantically well described with an explicit mapping property. Therefore, there is an obvious need

to make these mappings available to the community in a repository, avoiding external users the burden of extracting them ontology per ontology.

We have extracted the declared mappings from the source files using OMHT[6], developed as a standalone Java program that works with one ontology source file pulled out from an ontology repository. The standard properties used by OMHT to identify declared mappings inside a source file are the following: `owl:sameAs`, `oboInOwl:hasDbXref`, SKOS mapping properties and optionally `rdfs:seeAlso`. OMHT processes semantic resources in XML/RDF syntax an relies on the ontology repository to deal with different representation languages. OMHT takes as input a set of AgroPortal ontology acronyms and returns a JSON file for each input ontology that stores extracted mappings along with their metadata. Sometime, the target ontology and term are not explicit (especially with OBO XRefs which do not use URIs) therefore, OMHT relies on a manually curated file to resolve ambiguous targets.

In the dataset, we have removed extracted mappings for which source and target ontology are the same e.g., AFO contains 421 `oboInOwl:hasDbXRef` mappings to concepts in the same ontology; similarly, PO contains 40 internal XRefs. Surprisingly, this situation happens quite often: we have found a total of 2,230 such internal mappings all of them using the `oboInOwl:hasDbXef` property. The use of `oboInOwl:hasDbXRef` for representing ontology mappings is controversial as this property is used in the OBO community to capture several pieces of information including mappings between ontologies e.g., cross-references to database or database entries, curators of terms, references to publications, etc. In this study, we have carefully curated only the XRefs that correspond to ontology mappings (11% of them) to build our corpus as explained in another publication [13]. For instance, we have excluded XRefs to URLs or databases. Finally, we distinguish internal, inter-portal, and external mappings respectively if the target ontology is in AgroPortal, another repository of the OntoPortal family or simply identified by its URI.

3.4 Final Mapping Dataset

The total number of mappings of this dataset is 444,496 as described by Figure 1 (left). Term reuse and term overlap mappings represent (59,2%) of the total number of mappings, whereas explicit usage of mapping properties inside the ontology source files represent 40,8%.[7] Figure 1 (right) represents the overlap between the three mappings constructs. This diagram shows also the number of unique mappings for each mapping construct. We found two sets of 1,278 and 49,563 of overlapping mappings, which represent 11,43% of the dataset. The first intersection is an uncommon and odd situation where ontology developers have declared an explicit mapping to a class being explicitly reused. The second intersection is more interesting: it shows how much the number of lexical match in our corpus are explicitly identified as declared mappings by ontology developers. One would like to see this intersection grows.

[6] `https://github.com/agroportal/ontology_mapping_harvester`
[7] Our mapping dataset is publicly available at `https://bit.ly/3gFJ2DD`.

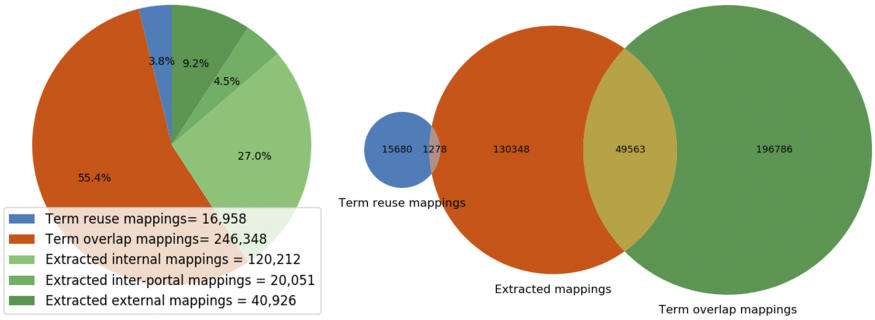

Fig. 1. Number of mappings (left). Venn diagram of the mapping dataset (right).

4 Mapping Dataset Analysis

Our goal in this study is to investigate the occurrence of patterns from the collected mappings. Therefore, we have built several mapping graphs that represent semantic resources and their alignments (i.e., set of mappings) respectively as nodes and edges. We can visualize these graphs based on the percentage of alignment, as described hereafter and provide an individual and a combined visualization for each mapping construct. Thus, we can identify hubs and clusters of semantic resources in our dataset. We expect such visualization will help ontology developers to better understand the ontology landscape in their domain of interest and possibly improve their semantic resources.

Similar to Ghazvinian et al. [7] percent-normalized link, we compute the percentage of mappings $\mathcal{P}$ by dividing the number of mappings $\mathcal{M}$ between a pair of ontologies $\mathcal{O}_s$ and $\mathcal{O}_t$ by the total number of concepts $|\mathcal{V}_s|$ of the source ontology based on the following formula: $\mathcal{P} = |\mathcal{M}(\mathcal{O}_s, \mathcal{O}_t)| \ / \ |\mathcal{V}_s|$. For instance, if an ontology $\mathcal{O}_1$ has 1000 terms, and 500 of these terms are mapped to terms in an ontology $\mathcal{O}_2$, then $\mathcal{P}(\mathcal{O}_1\ \mathcal{O}_2) = 50\%$. If one ontology is much larger than another, a large fraction of the small ontology may be mapped to the large one, but the set of mappings still constitutes a small percentage of the large ontology. This formula helps to investigate the level of mappings compared to the size of source ontologies.

4.1 Term Reuse Analysis

Out of a total number of 3,725,495 declared classes or concepts in our corpus, we found 16,958 term reuse mappings, with an average percentage $\mathcal{P}$ of 18,28% between pairs of semantic resources where at least one URI was explicitly shared between at least a pair of semantic resources. Out of 109 AgroPortal resources, 39 do not reuse any term from another ontology in the corpus which means that 70 does; but the number of distinct pairs of semantic resources is 174, which is quite low. The percentage of reuse is mostly under 5%, however, we found 42 pairs of ontologies with a term reuse above 10% and 8 pairs exhibit term reuse

between 95% and 100% which illustrates a situation where an ontology almost completely reuse another one.

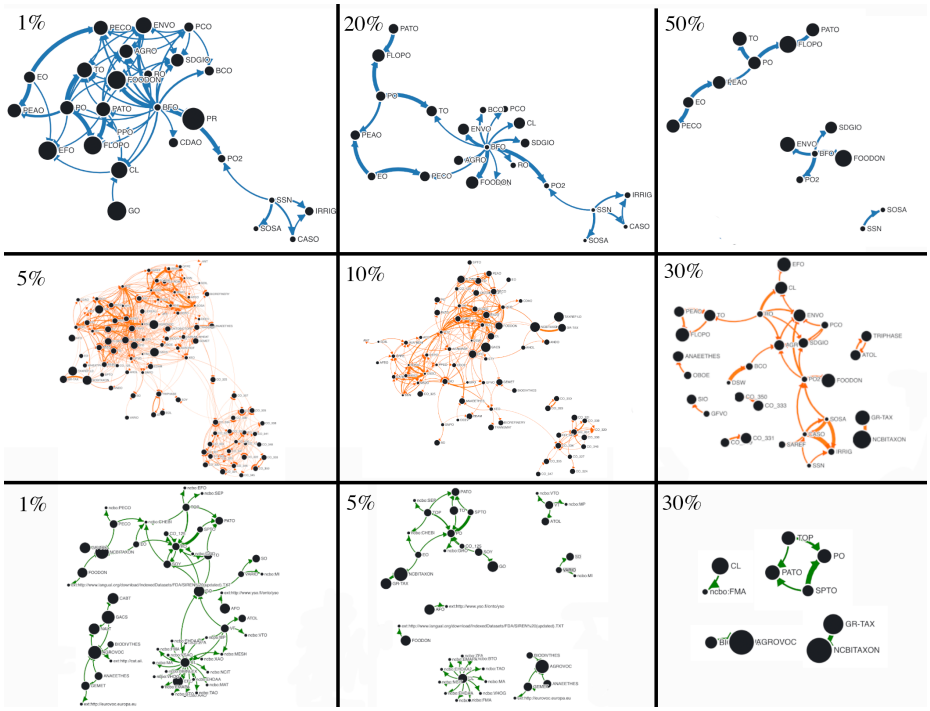

Fig. 2. The three mappings construct with different values of $\mathcal{P}$, arrows read as "is mapped to". Thickness of nodes and edges are respectively proportional to the sizes of the semantic resources and the percentage of mappings between them. Row 1: term reuse; Row 2: term overlap; Row 3: extracted mappings.

Figure 2's raw 1 represents the term reuse graphs at different percentages. In other words, we display an arrow between a pair of semantic resources, if at least P% of the source is reused in the target e.g., terms from BFO are being reused within ENVO. We can conclude from Figure 2: **(1)** In the family of ontologies relying on the Basic Formal Ontology (BFO) upper level ontology –a total of 17 in our corpus– we distinguish important differences in the degree of reuses: from 2,77% (for CDAO) to 97,77% (for FOODON). **(2)** Some ontologies within the same area or build by the same group highly reuse one another e.g., the Plant Ontology (PO), Plant Trait Ontology (TO), Plant Experimental Conditions Ontology (PECO) and Plant Environment Ontology (EO) all developed in the Planteome project form a cluster. **(3)** Some ontologies are mostly built from reuse e.g., PECO reuses all the URIs of EO. **(4)** We can visualize different clusters often built around reference upper level ontologies (BFO) or

reference standards (SSN or PO). Different values of $\mathcal{P}$ lead to different clusters. For instance, at $\mathcal{P}$=20%, we can distinguish a cluster around the SSN ontology being reused by a number of ontologies such as CASO and SOSA. **(5)** We only find term reuses in OBO and OWL ontologies. These ontologies tend to reuse URIs from each other as encouraged for instance by the OBO principles. However SKOS vocabularies or reference thesauri tend to systematically declare their own URIs and use mapping properties to align with the other ontologies.

4.2 Term Overlap Analysis

We found a total of 246,348 term overlap mappings between the 109 semantic resources of our corpus. With related ontologies, a small number of term overlap is very common: 6,204 pairs have at least one term overlap and only 12 semantic resources did not have any term overlap with any other ones in the corpus. Therefore, there is good lexical similarity in our dataset, even if the majority of the pairs contains less than 5% of term overlap with an average of 2.05%. We found 98 pair of semantic resources having a term overlap percentage more than 10%. Figure 2's raw 2 represents the term overlap graphs at different percentages. For example, there is 51 806 term overlap between the Gramene Taxonomy (GR-TAX) and NCBI Taxonomy. Figure 2 reveals other practices and information about the ontologies in our corpus: **(1)** Some resources strongly overlap with other related ones, without explicitly using mappings properties e.g., TRIPHASE and ATOL. **(2)** Some resources are definitively about the same area but have nothing to do one another with respect to community of developers, common practices, or funding project. For instance, we can visualize at $\mathcal{P}$= 20% that BFO and the Semanticscience Integrated Ontology (SIO) are two upper level ontologies developed for different purposes and by different communities but unsurprisingly contains a certain level of overlap. The same observation can be made for the Biological Collections Ontology (BCO) and the Darwin Core vocabulary (DSW) which are two resources developed to facilitate biodiversity data interoperability. **(3)** Term overlap allow us to discover cases where a thesaurus relies on an upper level ontology but without explicitly reusing its objects. For instance, the ANAEE Thesaurus's design is inspired from OBOE but the thesaurus being exclusively developed in SKOS cannot explicitly relies on OBOE developed in OWL. We observe 38% term overlap between them. **(4)** At $\mathcal{P}$=30%, we visualize several clusters e.g., between FLOPO, TO and PEAO. This cluster is visible for both term overlap and term reuse but through the PO hub in the case of term reuse. Despite these strong connections, we will see after, that we do not find any usage of mapping properties between these ontologies.

4.3 Extracted Mappings Analysis

Out of 109 semantic resources, we found 81 (74,31%) do not declare any mappings. From the 28 other resources, we have extracted 181,189 mappings from source files and found 174 pairs. Figure 1 (left) shows the majority of extracted mappings are internal i.e., between AgroPortal resources, which tends to corroborate the thematic coherence of the repository. 11% of these mappings pointing to

target semantic resources in the NCBO BioPortal reveals the thematic proximity with biology and life sciences (e.g., environment, nutrition). Among the important targets in the NCBO BioPortal are the Foundational Model of Anatomy (FMA) with 3,431 mappings or the ChEBI ontology with 745 mappings from 11 ontologies. External mappings target semantic resources that are not yet hosted in an ontology repository which denotes: (i) the willingness of ontology developers to map to semantic resources beyond the original domain captured within an ontology –this is a good practice for linked open data; (ii) integration of semantic resources in domain-specific repositories is not over. Among the most important external targets, we can cite 20,699 mappings from AGROVOC to the Chinese Agricultural Thesaurus (CAT) not yet integrated in AgroPortal.

Figure 2 's raw 3 reveals practices and information about extracted mappings in our corpus: **(1)** Every important reference thesaurus in AgroPortal (AGROVOC, ANNAETHES, NALT, GEMET) is strictly aligned to other ones in the domain, which seems to be a better practice than for ontologies in the wild. **(2)** Some semantic resources lexically very close (term overlap) are also formally aligned, like the case of GR-TAX being aligned to NCBITAXON. Indeed, when designed GR-TAX employed a lot of terms from NCBITAXON but the developers have decided to create new URIs and declared mappings between them. **(3)** At different levels of $\mathcal{P}$, we visualize some clusters different from the ones observed before e.g., around PO, a cluster is formed with different ontologies such as the TOP thesaurus which is developed by a different project. **(4)** We can observe a surprisingly low count of `owl:sameAs` in our dataset (3,255/181,189). Whereas this property was originally proposed explicitly for mappings, its strong logic entailment results in ontology developers not using it at the benefit of SKOS properties that do not have any logical entailment.

4.4 Combined Mappings Visualization and Analysis

Using an interactive visualization, we can see links between semantic resources for any mapping constructs and identify prominent hubs and clusters with variation of $\mathcal{P}$. It is available online with the ObservableHQ Web application: `https://observablehq.com/@amirlad?tab=collections`. Interested users can visualize each mapping construct individually and combined and dynamically change the percentage threshold. We believe, such visualization could be useful to ontology developers to select semantic resources for reuse or alignment.

Figure 3 (right) shows two hubs identified in our dataset: (i) PO, with mappings from and to 10 semantic resources; (ii) BFO, with terms being reused by many ontologies. Based on the combined visualization in Figure 3 (left), we can also visualize other prominent hubs. NCBITAXON, with a set of 59,186 mappings coming from 6 other semantic resources (PECO, TO, CL, FOODON, GR-TAX, EO) counts for 47% of the total number of internal mappings in the dataset. Figure 3 (left) depicts a combined graph at different values of $\mathcal{P}$ for each mapping construct. In this graph, we easily visualize several clusters and how they involve different constructs. For instance, we can visualize a 5-resource cluster (SSN, SOSA, CASO, IRRIG, SAREF) in which a mix of term reuse

and term overlap mappings interconnect the ontologies but no explicit declared mappings. We can also visualize a cluster of ontologies around AGROVOC.

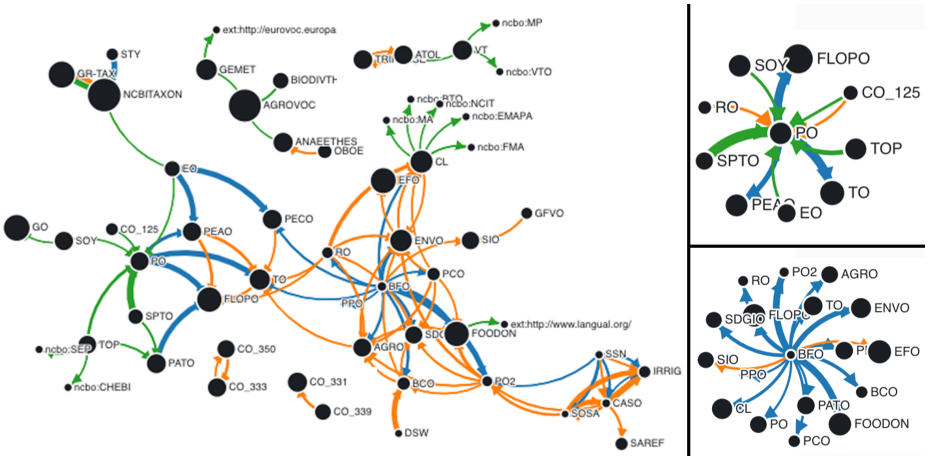

Fig. 3. (left) Combined graph of the three mapping constructs with $\mathcal{P}$= 20% for term reuse, $\mathcal{P}$= 35% for term overlap and $\mathcal{P}$= 10% for extracted mappings. (right) Plant Ontology and Basic Formal Ontology hubs for all constructs at $\mathcal{P}$=15%.

5 Discussion

Observations specific to ontologies in the OBO Foundry. For the 22 OBO Foundry ontologies in our corpus, there are 13,340 term reuses, 45,907 term overlaps mappings, and 2,799 extracted mappings. We found an average term reuse percentage of 15,06% between 73 pairs of OBO Foundry ontologies. Indeed, favoring term reuse is one principle of the OBO Foundry community. We found an average term overlap percentage of 4,92% between 345 pairs of OBO Foundry ontologies. In complement, we find an average percentage of extracted mappings of 1,73% between 48 pairs of OBO Foundry ontologies. Only two pairs of ontologies exhibit a percentage over 10 %: EO-PO and EO-NCBITAXON with respectively a percentage of 15,65% and 21,53%.

Observations specific to the Crop Ontology project. The Crop Ontology project [14] counts 22 ontologies in our corpus. We did not find any term reuse or extracted mappings between them. However, these ontologies often rely on a set of terms, especially for upper-level entities, described in a namespace that corresponds to no ontology e.g., `http://www.cropontology.org/rdf/Measurement`. With the mapping visualization, at different levels of $\mathcal{P}$, we can observe a term overlap cluster made by several of the ontologies in the Crop Ontology projects. These ontologies define similar traits of different crops, so it is normal they display a strong overlap. We found a total of 31,368 term overlap mappings between 640 pairs with an average percentage of 3,85%.

The term overlap percentage for the crop ontologies is slightly higher than the term overlap percentage in the corpus. However, the crop ontologies are not well reused or mapped by other ontologies in AgroPortal.

Observations specific to SKOS thesauri. Over 10 SKOS thesauri in our corpus, we did not find any term reuse mappings. We found 1,792 term overlap mappings and 41,932 of extracted mappings which tend to say the thesauri do not strongly overlap, even if they are well aligned with one another. There is an average percentage of 1.05% of term overlap between 52 pairs of the 10 SKOS thesauri of AgroPortal. We only found the use of mapping properties between 8 pairs with an average percentage of 10,35%. Reference thesauri developed by large organizations (e.g., FAO, USDA) do not reuse URIs from other semantic resources even if they overlap. But, they tend to declare explicit mappings using the SKOS mappings property more than the rest of the semantic resources in the corpus. For instance, AGROVOC, which is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization, do not reuse any terms from other semantic resources; however, it is explicitly aligned to GEMET and NALT. Thesauri tend to develop their URIs rather than reusing other URIs then aligning the copied terms to the original thesaurus. Unlike OBO Foundry ontologies, there is a lack of collaborative effort to develop SKOS thesauri that employ the same terms. But, when mappings are explicitly declared, they are well encoded and fully reusable as not in the XRefs, which are semantically ambiguous and need to be curated.

To the best of our knowledge, there is no other analysis of mappings in the domain covered by AgroPortal. However, in the following, we compare our results to the three most relevant mapping analysis studies identified in the related work.

Analogy with Ghazvinian et al. 2009 [7]. They analyzed a set of 4 million term overlap mappings for 207 biomedical ontologies stored in BioPortal and UMLS. Their dataset contained more than 4 million concepts. Here, we studied term reuse mappings, term overlap mappings, and extracted mappings from 3,735,344 concepts of 109 ontologies stored in AgroPortal. The total number of mappings is 444,496 with 246,348 term overlap mappings. We can deduce there is less term overlap in agri-food ontologies than biomedical ontologies. Indeed, Ghazvinian et al. reported that biomedical ontologies are very closely connected, with 33% of them having at least half of their concepts mapped to concepts in other ontologies. Whereas, in our dataset only 20 ontologies (18,34%) have at least 50% of their terms mapped to terms in some other ontologies. Therefore, there is less term overlap in our agronomy and biodiversity dataset than in the biomedicine dataset. Ghazvinian et al. stated that in biomedicine there is a little bit of overlap in everything, resulting in the extremely connected graph at $\mathcal{P}$=1%. At $\mathcal{P}$=20%, however, they report a meaningful power-law distribution. In our corpus, we visualize a similar observation for the term overlap mappings construct. With 2268 ontology pairs at $\mathcal{P}$=1% and 132 ontology pairs at $\mathcal{P}$=20%. However, for term reuse mappings and extracted mappings, the power-law distribution is lower than for term overlap mappings. For term reuse mappings, we found 61 ontology pairs at $\mathcal{P}$=1% and18 ontology pairs at $\mathcal{P}$=20%. Dealing with

extracted mappings, we found 68 ontology pairs at $\mathcal{P}$=1%, and 18 at $\mathcal{P}$=20%. Ghazvinian et al. visualized only term overlap mappings, however in our case, we can generate a combined visualization of the three mapping constructs. They stated term overlap mappings can be employed to identify prominent ontologies in a domain. This is true in our study too plus, the combination of the mapping constructs helps to have a better overview of the existing prominent ontologies.

Analogy with Poveda et al. 2012 [18]. They reported a percentage of 40% of term reuse in 196 semantic resources in the Linked Open Vocabularies (LOV) registry, which do not contain agri-food or biodiversity-ecology semantic resources. This percentage is higher than the average percentage (18,28%) of term reuse in AgroPortal ontologies.

Analogy with Kamdar et al. 2015 and 2017 [11, 12]. They first reported an average percentage of term overlap of 14.4% across 377 biomedical ontologies then in 2017, they reported a higher term overlap percentage (22.23%). For 109 AgroPortal ontologies, we found an average term overlap percentage of 2.05%, which is much lower than reported for BioPortal. This is mostly due to the method used to find lexical similarity. With a lower threshold and with the removal of stopwords, Kamdar et al.'s method keeps more term overlap mappings than LOOM (higher recall). However, our method can result in a better precision, even if we acknowledge it has certain limitation: lexically-similar labels in different ontologies may represent totally different concepts. Kamdar et al. considered XRefs as term reuse mappings, however, they do not consider other mapping properties. In our study, we extracted mappings with all the mapping properties available in the ontologies (including XRefs) and kept for term reuse only entities using the same URIs. This approach allows us to derive better insights from our dataset. Similarly to Kamdar et al., we found that most ontologies reuse less than 5% of their terms. This is contrary to the orthogonality principle encouraged in ontology engineering [1].

6 Conclusions and Perspective

We have built and analyzed a dataset of three mapping constructs based on a corpus of 109 semantic resources from AgroPortal. We have gathered more than 400,000 mappings either generated from AgroPortal or contained in the ontology source files. Our finding shows that most ontologies overlap with, reuse, or map less than 5% of their terms to other ontologies. Some communities have adopted certain good practices that it would be valuable to share with others. For instance, term reuse is more common in ontologies from the OBO Foundry, however the way these ontologies encode declared mappings is bad. On the other hand, term reuse is nonexistent in reference to SKOS thesauri, however these thesauri have a clear and consistent use of SKOS mapping properties for their declared mappings.

Despite the recent promotion of the FAIR data principles [21], which apply to semantic resources as any other data, some efforts are still necessary to interconnect them. Overall, ontology developers sometimes copy terms from other

semantic resources or define terms without checking reusable ontologies –which result in term overlap– or without explicitly reusing them or explicitly mapping them to the source ontology. Coming back to Figure 1 (right), a better situation would be to have a blue circle (term reuse) as big as possible, which would consequently decrease the size of the orange circle (extracted mappings) in which most of the green circle would be included (term overlap) making the yellow intersection (overlap with explicitly declared mappings) much of it.

The main contribution of our paper is the analysis and the visualization of these three mapping constructs which we hope will serve ontology developers to improve their practices and build semantic resources that will be as much as possible interoperable, reusable, and reused. This analysis can lead to relevant insights on the characteristics of the mappings repository. Since the use of ontologies, thesauri, and taxonomies expands, this visualization and its analysis can play an important role in understanding the relationships between semantic resources, and to identify clusters and hubs. We hope that these findings will be used to develop better guidelines, enhance term reuse and the use of mappings properties, and minimize term overlap.

The number of ontologies in AgroPortal increase and they are constantly updated. As future work, we plan to automate the analysis and visualization of term reuse, term overlap, and extracted mappings directly in AgroPortal. So that the subsequent version of the dataset used in this study could be automatically produced and exported from AgroPortal. We also plan to include an analysis and visualization of mappings for each ontology in the repository, which means that a developer will have an analysis, specific to his/her ontology. We are currently working on a new ontology alignment framework inside AgroPortal. This framework will contain a revised version of the ontology repository which shall generate term overlap mappings, identify term reuse mappings, extract declared mappings and also use external automatic matching systems to generate new mappings. Then each source of mappings will be merged into a unique alignment where each merged mappings will be scored and described with provenance.

Acknowledgements

This work was achieved with support of the AGRO Labex (ANR-10-LABX-0001), the NUMEV Labex (ANR-10-LABX-20) and the Data to Knowledge in Agronomy and Biodiversity (D2KAB – `www.d2kab.org`) project that received funding from the French National Research Agency (ANR-18-CE23-0017).

References

1. Amir Ghazvinian, Natalya F. Noy, M.A.M.: How orthogonal are the OBO Foundry ontologies? Biomedical Semantics **2**(2) (May 2011)
2. Annane, A., Bellahsene, Z., Azouaou, F., Jonquet, C.: Building an effective and efficient background knowledge resource to enhance ontology matching. Web Semantics **51**, 51–68 (August 2018)

3. Euzenat, J., Shvaiko, P.: Ontology matching, Second edition. Springer (2013)
4. Faria, D., Jiménez-Ruiz, E., Pesquita, C., Santos, E., Couto, F.M.: Towards Annotating Potential Incoherences in BioPortal Mappings. In: 13th Int. Semantic Web Conf., ISWC'14. LNCS, vol. 8797, pp. 17–32. Riva del Garda, Italy (Oct 2014)
5. Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating Mappings For Ontologies in Biomedicine: Simple Methods Work. In: American Medical Informatics Association Annual Symposium, AMIA'09. pp. 198–202. Washington, USA (Nov 2009)
6. Ghazvinian, A., Noy, N.F., Musen, M.A.: From mappings to modules: using mappings to identify domain-specific modules in large ontologies. In: 6th Int. Conf. on Knowledge Capture, K-CAP'11. pp. 33–40. Banff, Canada (June 2011)
7. Ghazvinian, A., Noy, N.F., Jonquet, C., Shah, N.H., Musen, M.A.: What Four Million Mappings Can Tell You about Two Hundred Ontologies. In: 8th Int. Semantic Web Conf., ISWC'09. LNCS, vol. 5823, pp. 229–242. Washington, USA (Nov 2009)
8. Jonquet, C.: Ontology Repository and Ontology-Based Services – Challenges, contributions and applications to biomedicine & agronomy. HDR Manuscript, University of Montpellier, Montpellier, France (May 2019)
9. Jonquet, C., Toulet, A., et al.: AgroPortal: a vocabulary and ontology repository for agronomy. Computers and Electronics in Agriculture pp. 126–143 (Jan 2018)
10. Jupp, S., Liener, T., Sarntivijai, S., Vrousgou, O., Burdett, T., Parkinson, H.: OxO A Gravy of Ontology Mapping Extracts. In: 8th Int. Conf. on Biomedical Ontology, ICBO'17. vol. 2137, p. 2. Newcastle, UK (2017)
11. Kamdar, M.R., Tudorache, T., Musen, M.A.: Investigating term reuse and overlap in biomedical ontologies. 6th Int. Conf. on Biomedical Ontology, ICBO'15 (July 2015)
12. Kamdar, M.R., Tudorache, T., Musen, M.A.: A systematic analysis of term reuse and term overlap across biomedical ontologies. Semantic web **8**(6), 853–871 (2017)
13. Laadhar, A., Abrahao, E., Jonquet, C.: Investigating One Million XRefs in Thirthy Ontologies from the OBO World. In: 11th Int. Conf. on Biomedical Ontologies, ICBO'20. Bozen-Bolzano, Italy (Sept 2020)
14. Matteis, L., Chibon, P., et al.: Crop ontology: vocabulary for crop-related concepts. In: 1st Int. Work. on Semantics for Biodiversity. pp. 37–46. Montpellier, France (May 2013)
15. Ngo, D., Bellahsene, Z.: YAM++ : A Multi-strategy Based Approach for Ontology Matching Task. In: 18th Int. Conf. on Knowledge Engineering and Knowledge Management, EKAW'12. LNCS, vol. 7603, pp. 421–425. Galway, Irland (Oct 2012)
16. Noy, N.F., Shah, N.H., et al.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research pp. 170–173 (May 2009)
17. Pathak, J., Chute, C.G.: Debugging Mappings between Biomedical Ontologies: Preliminary Results from the NCBO BioPortal Mapping Repository . In: Int. Conf. on Biomedical Ontology. pp. 95–98. Buffalo, USA (July 2009)
18. Poveda Villalón, M., Suárez-Figueroa, M.C., Gómez-Pérez, A.: The landscape of ontology reuse in linked data. Ontology Engineering in a Data-driven World, OEDW'12) (Oct 2012)
19. Till, M., Kutz, O., Codescu, M.: Ontohub: A semantic repository for heterogeneous ontologies. In: Theory Day in Computer Science, DACS'14. p. 2. Bucharest, Romania (Sept 2014)
20. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. Semantic Web (2014)
21. Wilkinson, M.D., Dumontier, M., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3** (March 2016)