Cross-querying LOD datasets using complex alignments: an application to agronomic taxa

Elodie Thiéblin[1], Fabien Amarger[1], Nathalie Hernandez[1], Catherine Roussey[2], and Cassia Trojahn Dos Santos[1]

[1] IRIT UMR 5505, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9,
firstname.lastname@irit.fr
[2] Irstea, 9 avenue Blaise Pascal CS 20085 63178 Aubière,
firstname.lastname@irstea.fr

Abstract. Farmers have new information needs to change their agricultural practices. The Linked Open Data is a considerable source of knowledge, separated into several heterogeneous and complementary datasets. This paper presents a process to query LOD datasets from a known ontology using complex alignments. The approach was applied on AgronomicTaxon, a taxonomic classification ontology, to query Agrovoc and DBpedia.

Keywords: Query Rewriting, Complex Alignments, Agronomic Sources, Linked Open Data

1 Introduction

The Linked Open Data (LOD) is a considerable source of knowledge, divided into several heterogeneous and complementary datasets. Following LOD principles, a dataset should be linked to others datasets. There exist two different kinds of links. Direct links between two datasets using properties like *owl:sameAs* or *rdfs:seeAlso*. These properties establish correspondences between entities of distinct datasets [19]. Indirect links when two datasets reuse some existing ontologies[3]. That means that the two dataset schemas share some common entity types. These kinds of links are mainly used to browse the different sources or to retrieve corresponding entities. A retrieval system should consider that the query is based on predefined ontology: either all datasets share the same ontology either a selected ontology is used as a reference and all dataset schemas are linked to it. With existing approaches, end-users interactions with the system are needed.

Hence, the LOD has become a needful source of knowledge in many domains, such as in the agriculture domain. For example, due to climatic change and their willingness to improve environmental impacts, farmers and agronomists must rethink agricultural practices. To do so, farmers need to find information about plant or any living organism. They can be looking for new crops that are able to better support their pedoclimatic conditions. Farmers can also find in their plots

[3] ontologies are defined as semantic web data schema

unknown insects and may want to know if they are pests of their crops. Furthermore, they can find in their plots some unknown plants and want to know if they are weed plants or auxiliary plants for their crops. These kind of information can be extracted from scientific sources as presented in [10]. To answer to these information needs, farmers needs to query several datasets that describe living organisms. In such a domain, users are becoming more and more familiar with particular LOD datasets and are able to query them with SPARQL[4]. However, as many domain datasets are nowadays published on the LOD, complementary information may be relevant in other sources. Reformulating the information need according to other ontologies is time consuming. Ontology represents a specific point of view on the domain often influenced by the application needs. Exploiting available ontologies implies taking into account the different modelling issues of the same domain. We propose to take into account this aspect by considering complex correspondences between ontologies [22].

In this paper, we propose a method for helping end-users query the LOD when they have a specific need and have expressed it on a first dataset that can satisfy it. The main idea is to automatically reformulate their SPARQL query by using correspondences established between the different ontologies to find complementary information in the other datasets. The originality of our proposition is to take into account complex correspondences which define expressive correspondences between the ontologies. A first experiment has been carried out with agriculture domain experts that have specific needs dealing with agronomic taxa.

The paper is organised as follows. First we present the context of this work by describing available sources in the agronomic domains and existing Semantic Web approaches dealing with query reformulation. Then we give an overview on our approach. Finally we detail the results we obtained when applying our approach on agronomic taxa.

2 Context

2.1 Agricultural sources

For information needs related to living organisms, farmers may query several type of information sources available on the Web; For example NCBI[5], TaxRef[6] or Encyclopedia of life[7]. Unfortunatelly these sources are not represented in the Semantic Web formalisms which makes them difficult to query automatically. For that reason these sources are out of scope of this paper, we will focus only on sources available on the LOD.

[4] the W3C recommandation`https://www.w3.org/TR/sparql11-query/` for a query and update language for the Semantic Web.

[5] `http://www.ncbi.nlm.nih.gov/taxonomy`

[6] `https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref`

[7] `http://eol.org/`

AgronomicTaxon [8] When searching for information related to plants, AgronomicTaxon [18] can be considered as it is, as far as we know, the well formalised ontology available on the domain. This ontology has been developed using the NeOn methodology and reuses several sources and ontology design patterns.

As described in [18] this ontology models the taxonomy thanks to a central class which is *agro:Taxon*. This class is specialised in several sub-classes to model the different levels in the taxonomy. This specialisation, from *agro:VarietyRank* to *agro:KingdomRank*, uses the *agro:hasHigherRank* property to link the different levels. Each *agro:Taxon* is described with vernacular and/or scientific names.

To populate this ontology we used the Muskca system described in [1]. The Muskca system output has been validated manually. The final output deals with the wheat taxonomic classification. We chose this sub-domain to avoid a large number of concepts and allow the manual validation.

DBPedia [9][2] is a dataset based on the Wikipedia[10] data export.

This dataset covers a lot of domains and is populated with several millions of individuals. For this reason, it is largely used as an instance alignment reference: its instances are linked with other datasets' instances. DBpedia can be seen as a hub between different sources on the LOD. Nevertheless, the Wikipedia policy such as community participation and correction, brings some errors and approximation in its model. The part that deals with agronomic classification contains a large number of taxa but the model is unclear. For example the resource *dbo:Eukaryote* is defined as a *owl:Class*. The resource *dbr:Eukaryote* is defined as an individual. The redundancy here brings some ambiguities because some taxa have the relation *dbo:domain* with the individual *dbr:Eukaryote* and others are typed with the class *dbo:Eukaryote*.

Agrovoc [11][4] is maintained by experts all over the world thanks to the VocBench platform and overviewed by the Food and Agriculture Organization of the United Nations (FAO). Agrovoc covers all the FAO's areas of interest, such as agriculture, forestry, fisheries, food and related domains. It is available in 20 languages, with an average of 40,000 terms per language. AGROVOC is available in SKOS-XL[12] (with close to 32,000 concepts), and published as Linked Open Data.

The strength of this thesaurus is the lexical coverage with a large number of concepts. The second strength is that this source is used as a reference by experts who wants to manage agricultural data. These strength especially concern the agronomic taxonomic part. Nevertheless, this source cannot be manipulate easily because of its update policy. It is updated manually by some experts, then we can found some erroneous information and information lack [20]. Concerning the agronomic classification we can encounter some ambiguities because of the thesaurus model. The hierarchical relation (skos:broader/skos:narrower) can

[8] we will use the prefix *agro* for the reference of this ontology

[9] `http://dbpedia.org/` prefixes: *dbo* ($\mathcal{T}box$), *dbr* ($\mathcal{A}box$)

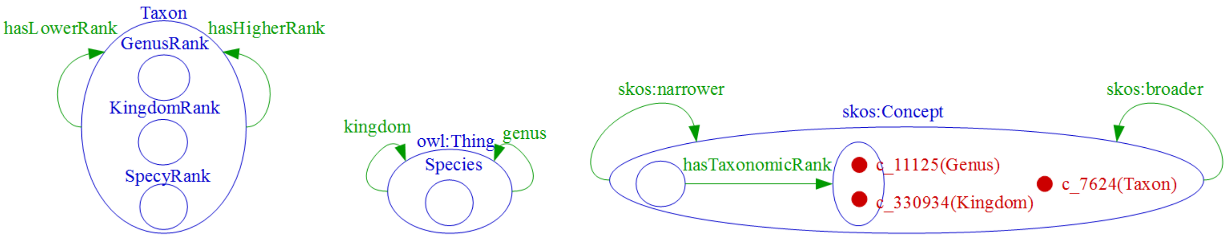
[10] `https://fr.wikipedia.org/`

[11] `http://aims.fao.org/aos/agrovoc/`, prefixes: *agronto* ($\mathcal{T}box$), *agrovoc* ($\mathcal{A}box$)

[12] `http://www.w3.org/TR/skos-reference/skos-xl.html`

represent several kind of relation (subsumption, partOf, domain specific specialisation). For example, the skos:Concepts link by the skos:narrower property to the skos:Concept ”Triticum” are ”Winter wheat” and ”Summer wheat”. They are not known to be scientific agronomic rank.

Figure 1 presents fragments of the 3 sources in [21]'s visualisation format.



(a) AgronomicTaxon (b) DBpedia (c) Agrovoc

Fig. 1: Fragments of the three ontologies in [21]'s visualisation format

2.2 Alignment definition and applications

As can be observed for the 3 fragments of the sources presented in the figure 1, the ontologies describing the datasets are heterogeneous. The purpose of ontology matching is to reduce the heterogeneity between ontologies [9]. Ontology matching is the process of generating an ontology alignment A between two ontologies [9]: a source ontology o and a target ontology o'. A is a set of correspondences. Each correspondence is a triple $\langle e_o, e_{o'}, r \rangle$

- if the correspondence c_i is **simple**, both e_o and $e_{o'}$ are atomic entities (e.g. class, object property, data property or instance). One IRI is matched with another IRI (1:1), e.g. $\forall x$, $\mathit{agro{:}Taxon}(x) \equiv \mathit{dbo{:}Species}(x)$ is a simple correspondence.
- if the correspondence is **complex**, at least one of e_o or $e_{o'}$ involves one or more atomic entities in a logical construction. The correspondence is therefore (1:n), (m:1) or (m:n) according to the number of entities, constructors or functions involved on each side of the correspondence.
  $\forall x$, *agro:GenusRank(x)* $\equiv$ *agronto:hasTaxonomicRank(x,agrovoc:c_11125)* is a (1:n) complex correspondence with more than one entity.
- r is a relation: equivalence ($\equiv$) or subsumption ($\geq$, $\leq$) between e_o and $e_{o'}$;

An alignment is called a *complex alignment* if at least one of its correspondences is complex. Approach that automatically generate complex alignments between ontologies are emerging [17,16,14]. Ontology alignments can be used for various applications such as ontology merging [24,15] and query mediation [9].

In this paper, the purpose is to query the agronomic sources on the LOD without transforming the ontologies or the data they contain. We consider this hypothesis because we want to preserve the different point of view available in each source. This is especially true for the agronomic classification domain because the domain experts do not agree on which classification they should use.

This is why query mediation is the most adapted solution. These methods use correspondences to rewrite queries to adapt to the second ontology. We need here to consider rewriting queries using complex correspondences because AgronomicTaxon/Agrovoc and AgronomicTaxon/DBpedia correspondences cannot be expressed with only simple correspondences. Some complex correspondence examples are presented in the section 4.1.

3 Related work

A SPARQL query is intrinsically related to the ontological model that describes the RDF source. To federate information from different sources described by various ontologies, a SPARQL query must be adapted to each of them.

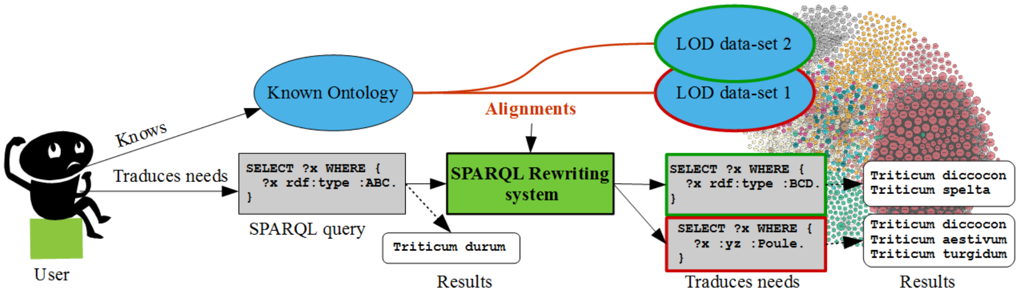
A naive approach for rewriting SPARQL queries consists in replacing the IRI of an entity of the initial query by the corresponding IRI in the alignment, using simple correspondences. This approach is integrated in the Alignment API [7]. However, it does not take into account the specific kind of relation expressed in the correspondence (e.g., generalisation or specialisation). Makris *et al.* [13,12] present the SPARQL-RW rewriting framework that applies a set of predefined rules for (complex) correspondences. They define a set of correspondence types on which the rewriting process is based (i.e., *Class Expression*, *Object Property Expression*, *Datatype Property*, and *Individual*). Zheng *et al.* [25] propose a rewriting algorithm that serves the purpose of context (i.e, units of measure) interchange for interoperability. Correndo *et al.* [5] apply a declarative formalism for expressing alignments between RDF graphs to rewrite SPARQL queries. In [6], a subset of EDOAL expressions are transformed into a set of rewriting rules. The expressions involving the restrictions on classes and properties and the restrictions on property occurrences and values are not featured in the rewriting rules. Thiéblin *et al.* [22] propose a set of rewriting rules from EDOAL expression. In comparison with Correndo *et al.*'s [6] approach, it can deal with expressions such as Class by Attribute Occurrence. The approach is based on the assumption that the queries to be transformed aim at retrieving new instances to meet a given need. This is why only $\mathcal{T}box$ elements are taken into account.

Some approaches have also been proposed in order to query several LOD datasets, thus helping the users to adapt the expression of their need to several sources. [23], for example, relies on explicit correspondences expressed within the dataset (with *owl:sameAs*, *owl:equivalentClass*, or *owl:equivalentProperty* properties) to automatically reformulate queries. Another example come from the SemaGrow[11] project. One use case is about querying multiple bibliographic datasets related to agricultural domain. All the dataset schemas share the same ontology. Some simple alignement between individuals are used to translate the query into the targetted dataset. The performance of this kind of methods depends on how datasets are explicitly linked. Most of the time, only simple correspondences between instances (expressed with *owl:sameAs* property) are expressed which limits the possibility of reformulation. Instances typed by classes or linked by properties for which no correspondences have been established will not be retrieved. The approach presented in [3] helps end-users express their query by means of a graphical interface that automatically adapts to a specific selected LOD dataset. An overview of the ontology used to describe the data is presented and the interface assists the formulation of the query according to it. More intuitive, SimplePARQL[8] proposes a way for formulating SPARQL queries by using terms for designating resources instead of their IRI. The users do not need to know the underlying ontology but this approach implies that the ontology is exhaustively annotated with all the *rdfs:label* that can be associated with the resources. The aim of our work is to reformulate queries automatically by using the expressiveness of complex correspondences.

4 An approach for querying LOD datasets

Figure 2 presents the global work-flow of the approach. The users knows an ontology (e.g. AgronomicTaxon) and can write a SPARQL query expressing their needs using this ontology. An ontology alignment exists between the known ontology and ontologies from the LOD (e.g. DBpedia, Agrovoc). The SPARQL rewriting system rewrites the query to query the LOD dataset. The users get the information fitting their needs from various sources. The SPARQL rewriting system as well as the alignments are publicly available[13].

The approach is illustrated by the use case with the known ontology being AgronomicTaxon, and two LOD datasets being Agrovoc and DBpedia.



```
SELECT ?x WHERE {
  ?x rdf:type :ABC.
}
```

```
Triticum durum
```

```
SELECT ?x WHERE {
  ?x rdf:type :BCD.
}
```

```
Triticum diccocon
Triticum spelta
```

```
SELECT ?x WHERE {
  ?x :yz :Poule.
}
```

```
Triticum diccocon
Triticum aestivum
Triticum turgidum
```

Fig. 2: Global work-flow of the approach

4.1 Ontology alignments

Our SPARQL rewriting system needs alignments between the known ontology and the LOD ontologies. Alignments between ontologies are not always available. In this case, some approaches may be able to generate them [17,14,16]. If the approaches are not exhaustive enough, the alignments can be manually written.

For our use case considering AgronomicTaxon, DBpedia and Agrovoc, these approaches did not generate any correspondence. For this reason, the alignments between the chosen ontologies were manually written in EDOAL[14] to apply the approach. AgronomicTaxon being the known ontology, it is the source ontology of the alignments. Each entity of AgronomicTaxon was put in correspondence

[13] https://framagit.org/IRIT_UT2J/sparql-translator-complex-alignment

[14] http://alignapi.gforge.inria.fr/edoal.html

when possible. During the alignment establishment phase, (1:n) correspondences were sought, equivalence correspondences were favoured over subsumption correspondences, simple equivalence correspondences were favoured over complex. This way, the correspondences are as simple and correct as possible. Table 1 presents a few correspondences of the alignment. The AgronomicTaxon-Agrovoc alignment contains 31 correspondences and the AgronomicTaxon-DBpedia alignment contains only 29 as 2 properties could not be translated. The granularity heterogeneity of the ontologies made it impossible for some entities (classes, object properties, data properties) to be put in an equivalence relation. For this reason, some correspondences have a subsumption relation. In correspondences (4) and (6), the $+$ symbol stands for the transitivity of the object property.

AgronomicTaxon entity	rel	Right member	Ref
$\forall x,\ agro{:}Taxon(x)$	$\equiv$	$dbo{:}Species(x)$	(1)
	$\equiv$	$\exists y,\ agronto{:}hasTaxonomicRank(x,y) \land skos{:}broader(y,agrovoc{:}c\_7624)$	(2)
$\forall x,y,\ agro{:}hasHigherRank(x,y)$	$\geq$	$dbo{:}Species(x) \land dbo{:}Species(y) \land\ (dbo{:}genus(x,y) \lor dbo{:}family(x,y) \lor dbo{:}order(x,y) \lor dbo{:}classis(x,y) \lor dbo{:}phylum(x,y) \lor dbo{:}kingdom(x,y))$	(3)
	$\leq$	$skos{:}broader{+}(x,y)$	(4)
$\forall x,y,\ agro{:}hasLowerRank(x,y)$	$\geq$	$dbo{:}Species(x) \land dbo{:}Species(y) \land\ (dbo{:}genus(y,x) \lor dbo{:}family(y,x) \lor dbo{:}order(y,x) \lor dbo{:}classis(y,x) \lor dbo{:}phylum(y,x) \lor dbo{:}kingdom(y,x))$	(5)
	$\leq$	$skos{:}narrower{+}(x,y)$	(6)
$\forall x,y,\ agro{:}prefScientificName(x,y)$	$\leq$	$rdfs{:}label(x,y)$	(7)
	$\leq$	$\exists\ z,\ skos{:}prefLabel(x,y) \lor (skosxl{:}prefLabel(x,z) \land skosxl{:}literalForm(z,y))$	(8)
$\forall x,\ agro{:}SpecyRank(x)$	$\leq$	$\exists y,\ dbo{:}Species(x) \land dbo{:}genus(x,y) \land dbo{:}Species(y)$	(9)
	$\equiv$	$agronto{:}hasTaxonomicRank(x,agrovoc{:}c\_331243)$	(10)

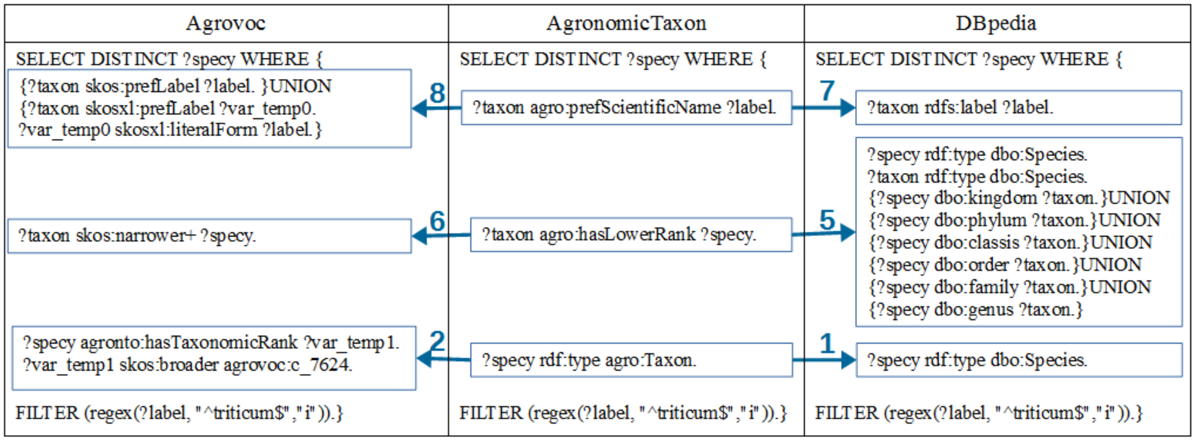
Table 1: Extract of correspondences of the AgronomicTaxon-DBpedia and AgronomicTaxon-Agrovoc alignments

4.2 SPARQL query rewriting from complex alignments

We use the rewriting SPARQL approach from [22] because it deals with complex alignments expressed in the EDOAL format and can process expressions such as *ClassByAttributeOccurrence* not processed by other systems. This system translates the triples of a SPARQL query one after the other. It can only translate two kinds of triples: Class Triples of the form `?x rdf:type o:SomeClass.` and Predicate Triples `?x o:predicate ?y.`. The subject of the triples of the query to be rewritten must always be a variable, the object is either a Class URI (in a class triple), a variable or a literal (in a predicate triple).

The following example presents the rewriting process of a SPARQL query on AgronomicTaxon for retrieving every subtaxa of Triticum. Table 2 presents the initial SPARQL query on AgronomicTaxon.The query is rewritten for Agrovoc and DBpedia using the alignments.

The first triple of the query is rewritten using the correspondences (8) for Agrovoc and (7) for DBpedia. The second triple was rewritten using (6) and (5).



Agrovoc	AgronomicTaxon	DBpedia
SELECT DISTINCT ?specy WHERE {	SELECT DISTINCT ?specy WHERE {	SELECT DISTINCT ?specy WHERE {
{?taxon skos:prefLabel ?label. }UNION {?taxon skosxl:prefLabel ?var_temp0. ?var_temp0 skosxl:literalForm ?label.}	?taxon agro:prefScientificName ?label.	?taxon rdfs:label ?label.
?taxon skos:narrower+ ?specy.	?taxon agro:hasLowerRank ?specy.	?specy rdf:type dbo:Species. ?taxon rdf:type dbo:Species. {?specy dbo:kingdom ?taxon.}UNION {?specy dbo:phylum ?taxon.}UNION {?specy dbo:classis ?taxon.}UNION {?specy dbo:order ?taxon.}UNION {?specy dbo:family ?taxon.}UNION {?specy dbo:genus ?taxon.}
?specy agronto:hasTaxonomicRank ?var_temp1. ?var_temp1 skos:broader agrovoc:c_7624.	?specy rdf:type agro:Taxon.	?specy rdf:type dbo:Species.
FILTER (regex(?label, "^triticum$","i")).}	FILTER (regex(?label, "^triticum$","i")).}	FILTER (regex(?label, "^triticum$","i")).}

Table 2: Original (AgronomicTaxon) and automatically rewritten SPARQL queries (Agrovoc and DBpedia) to retrieve sub-taxa of Triticum. The numbers are the correspondences references from table 1 used to translate each triple.

The last triple was rewritten using (2) and (1). The filter and the header are the same for all the queries. The query results on their respective dataset are shown in figure 3. The analysis of the results is detailed in the next section.

5 Result analysis : extracting information about AgronomicTaxon on the LOD

In this section, we present the information needs defined and we detail the results of the approach for each information need. The information needs considered in this experiment have been defined with domain experts when designing AgronomicTaxon. They are presented in table 3.

Question	Description
IN1	What is the rank of the taxon Triticum ?
IN2	What is the kingdom of the Triticum taxon ?
IN3	What are the common names of Triticum taxon in French ?
IN4	What are the common names of Triticum taxon in English ?
IN5	What are the different wheat species ?

Table 3: Information needs in natural language

Each information need was express with a SPARQL query for AgronomicTaxon. The results of the approach are presented below. Every SPARQL query and its results are available on the Framagit repository[15].

[15] https://framagit.org/IRIT_UT2J/sparql-translator-complex-alignment/tree/master/mtsr2017/

IN1: the rank of Triticum: Genus Rank The concept "Genus" is represented as a class in AgronomicTaxon and as an instance in Agrovoc and DBpedia. The SPARQL query on AgronomicTaxon specifies that the expected answer is a class with a *rdf:type* relation and uses the structure of the ontology through a *rdfs:subClassOf* relation. No answer is provided for the Agrovoc dataset as the rewriting approach can not properly translate the `?rank rdfs:subClassOf agro:Taxon` triple from the initial query using a complex correspondence. As correspondence (2) from Table 1 is a complex correspondence and should be used in the rewriting process, the triple is not well translated. For DBpedia, the (1) correspondence is a simple correspondence. Therefore, the triple can be translated. However, the structure of DBpedia is different from AgronomicTaxon's. The *dbo:Eukaryote* class is returned because Triticum is an instance of *dbo:Eukaryote* and *dbo:Eukaryote* is a subclass of *dbo:Species*. This answer is wrong and comes from the fact that some taxa in DBpedia are defined as classes (*dbo:Eukaryote*) and instances (*dbr:Eukaryote*).

This query used the structure of the source ontology which is very different from the target ontologies'. Therefore, the given results are poor.

IN2: the kingdom of Triticum: Plantae Plantae is an instance in AgronomicTaxon. Both rewritten queries are semantically correct. The query on Agrovoc retrieves the *agrovoc:c_330074* instance which is the Plantae taxon. The filter for the "wheat" label was changed to "triticum" because Agrovoc contains few vernacular names. The query on DBpedia does not retrieve anything because even though the *dbo:kingdom* property holds between *dbr:Triticum* and *dbr:Plantae*, *dbr:Plantae* was not specified as a taxon (*dbo:Species*).

This query was successfully rewritten for both target ontologies but DBpedia could not give an answer as some information is missing in the dataset.

IN3-4: vernacular names of Triticum in French and English: Blé and Wheat In AgronomicTaxon, the French vernacular name of Triticum is not specified while the English is. The AgronomicTaxon ontology distinguishes vernacular names from scientific names. In Agrovoc and DBpedia, no such distinction is made. Agrovoc uses *skos:prefLabel*, *skosxl:prefLabel* ∘ *skosxl:literalForm* properties to label its instances. The dataset mostly contains scientific labels. DBpedia uses the *rdfs:label* property and was populated with common names. In the alignments, the label properties of AgronomicTaxon could not find equivalents. The results of the queries are therefore more general than what is expected. Agrovoc returns "Triticum" as French common name and "Triticum" as English common name. DBpedia returns "Blé" for French and "Wheat" for English.

As there was no equivalence relation for the properties used in these queries but only more general properties, the results given by the queries is more general. The outcome of the query depends on the way the datasets were populated. The outcome of these queries shows the complementarity of the sources on the LOD. The "Blé" information was only present in DBpedia.

IN5: the species of Triticum genus Figure 3 presents the manual instance mappings made between the sub-taxa of Triticum in AgronomicTaxon and the instances in Agrovoc and DBpedia.

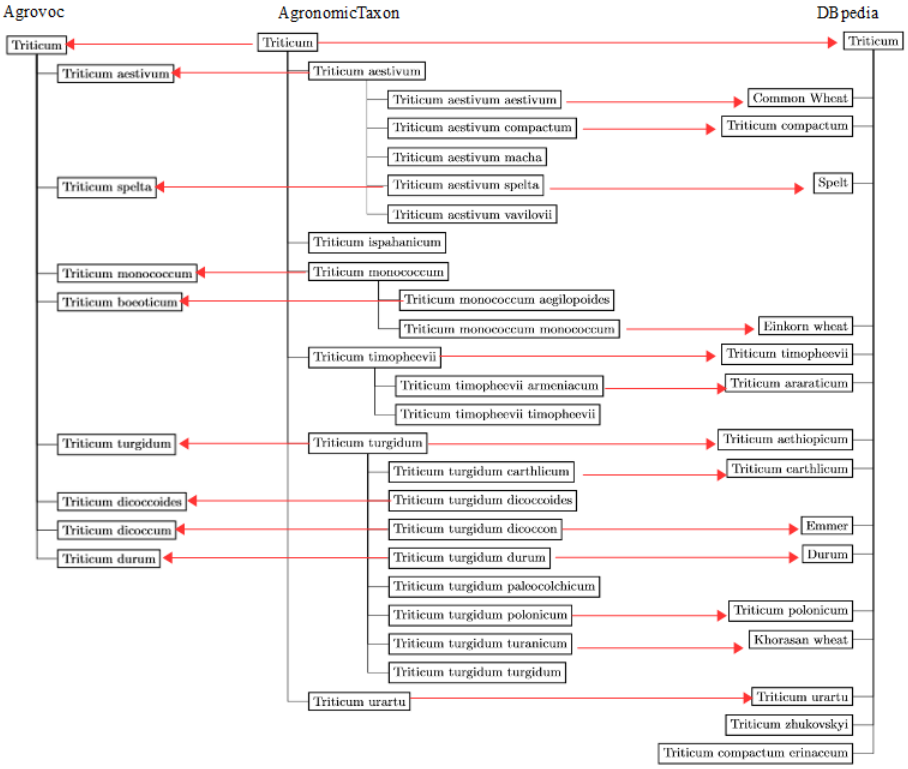


Fig. 3: Correspondences between the taxa under the Triticum genus in Agrovoc (left), AgronomicTaxon (center) and DBpedia (right). The red arrows are *rdfs:seeAlso* relations.

In the Agrovoc tree (left) a child relation between two nodes stands for a *skos:narrower* relation, in the AgronomicTaxon tree (left), it stands for a direct *agro:hasLowerRank* relation. In the DBpedia tree (right), a child has a *dbo:genus* relation with the *dbr:Triticum* taxon. In AgronomicTaxon (center), the species are the direct children of Triticum and the subspecies are the children of the species. In Agrovoc, all subtaxa of Triticum have a specie rank. There are no subspecies. In DBpedia, there is no rank distinction between the subtaxa of a genus. DBpedia is not as fine-grained as AgronomicTaxon so the class *agro:SpecyRank* has no equivalence in DBpedia. In the alignment, it is subsumed in correspondence (9) of table 1. Therefore the rewritten query will return all taxa below Triticum.

The result of this query emphasises the granularity heterogeneity between datasets. It also shows the complementarity of the sources as some sub-taxa of Triticum only appear in DBpedia and some only in AgronomicTaxon.

6 Conclusion

We presented the use of complex alignments for a SPARQL rewriting approach applied to agronomic LOD sources. This evaluation of this approach highlights a few points. First of all, the agronomic ontologies of the LOD are heterogeneous and simple alignments are not always expressive enough to provide interoperability between them. Secondly, LOD datasets contain complementary information based on how they are populated (IN3-4, fig. 3). A SPARQL rewriting approach based on complex alignments can show good results in information crossing depending on the nature of the query. Queries expecting a part of ontology as a result or using the structure of the query itself will not give good results with complex alignments (IN1). The granularity heterogeneity between ontologies will affect the semantic equivalence between two queries (IN5).This should be in mind when using the rewriting approach. Because of the scope heterogeneity between ontologies, some queries cannot be rewritten. The missing information of a dataset can affect the results of a query (e.g. DBpedia in IN2).

A few downsides can be stressed and shall be addressed in future works. First of all, the automatically obtained queries are not optimised for performance because of the triple by triple rewriting approach. This is an issue when querying large-scale datasets such as DBpedia. (m:n) correspondences are not processed by SPARQL query rewriting systems yet. A global interpretation of the SPARQL query, instead of a triple-by-triple process could be a solution to both problems. Another issue pointed out in IN1, is that the expected answer to a query can be a class in a dataset (*agro:GenusRank*) or an instance in another one (*agovoc:c_11125*). There still is no formalisation of this kind of correspondence (class-instance).

Acknowledgements

This work is partially supported by the French FUI SparkinData project.

References

1. Amarger, F., Chanet, J.P., Haemmerlé, O., Hernandez, N., Roussey, C.: Knowledge engineering method based on consensual knowledge and trust computation: The muscka system. In: International Conference on Conceptual Structures (2016)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. The semantic web (2007)
3. Benedetti, F., Bergamaschi, S., Po, L.: Lodex: A tool for visual querying linked open data. In: ISWC (2015)
4. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J.: The agrovoc linked dataset. Semantic Web 4(3), 341–348 (2013)
5. Correndo, G., Salvadores, M., Millard, I., Glaser, H., Shadbolt, N.: SPARQL Query Rewriting for Implementing Data Integration over Linked Data. In: 1st International Workshop on Data Semantics (DataSem 2010) (2010)
6. Correndo, G., Shadbolt, N.: Translating expressive ontology mappings into rewriting rules to implement query rewriting. In: 6th Workshop on Ontology Matching (2011)

7. David, J., Euzenat, J., Scharffe, F., Trojahn, C.: The Alignment API 4.0. Semantic Web 2(1) (2011)
8. Djebali, S., Raimbault, T.: Simpleparql: a new approach using keywords over sparql to query the web of data. In: Proceedings of the 11th International Conference on Semantic Systems. ACM (2015)
9. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer Berlin Heidelberg (2013)
10. Kulicki, P., Trypuz, R., Trójczak, R., Wierzbicki, J., Woźniak, A.: Ontology-based representation of scientific laws on beef production and consumption. In: Research Conference on Metadata and Semantic Research. pp. 430–439. Springer (2013)
11. Lokers, R., Konstantopoulos, S., Stellato, A., Knapen, R., Janssen, S.: Designing innovative linked open data and semantic technologies in agro-environmental modelling (2014)
12. Makris, K., Bikakis, N., Gioldasis, N., Christodoulakis, S.: SPARQL-RW: transparent query access over mapped RDF data sources. In: 15th International Conference on Extending Database Technology. ACM (2012)
13. Makris, K., Gioldasis, N., Bikakis, N., Christodoulakis, S.: Ontology Mapping and SPARQL Rewriting for Querying Federated RDF Data Sources. In: OTM Confederated International Conferences (2010)
14. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Discovering concept coverings in ontologies of linked data sources. In: ISWC. Springer (2012)
15. Pokharel, S., Sherif, M.A., Lehmann, J.: Ontology Based Data Access and Integration for Improving the Effectiveness of Farming in Nepal. IEEE (2014)
16. Qin, H., Dou, D., LePendu, P.: Discovering executable semantic mappings between ontologies. In: On the Move to Meaningful Internet Systems (2007)
17. Ritze, D., Völker, J., Meilicke, C., Sváb-Zamazal, O.: Linguistic analysis for complex ontology matching. In: 5th Workshop on Ontology Matching (2010)
18. Roussey, C., Chanet, J.P., Cellier, V., Amarger, F.: Agronomic taxon. In: Proceedings of the 2nd International Workshop on Open Data. p. 5. ACM (2013)
19. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: ISWC (2014)
20. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering thesauri for new applications: The AGROVOC example. Journal of Digital Information (2004)
21. Stapleton, G., Howse, J., Bonnington, A., Burton, J.: A vision for diagrammatic ontology engineering. In: International Workshop on Visualizations and User Interfaces for Knowledge Engineering and Linked Data Analytics (2014)
22. Thiéblin, É., Amarger, F., Haemmerlé, O., Hernandez, N., Trojahn, C.: Rewriting select sparql queries from 1: n complex correspondences. In: 11th Workshop on Ontology Matching (2016)
23. Torre-Bastida, A.I., Bermúdez, J., Illarramendi, A., Mena, E., González, M.: Query rewriting for an incremental search in heterogeneous linked data sources. In: International Conference on Flexible Query Answering Systems. Springer (2013)
24. Wang, Y., Wang, Y., Wang, J., Yuan, Y., Zhang, Z.: An ontology-based approach to integration of hilly citrus production knowledge. Computers and Electronics in Agriculture 113 (2015)
25. Zheng, X., Madnick, S.E., Li, X.: SPARQL Query Mediation over RDF Data Sources with Disparate Contexts. In: WWW Workshop on Linked Data on the Web (2012)