Finding Concept Coverings in Aligning Ontologies of Linked Data

Rahul Parundekar(✉), Craig A. Knoblock, and José Luis Ambite

Information Sciences Institute and Department of Computer Science,
University of Southern California, 4676 Admiralty Way, Suite 1001,
Marina del Rey, Sunnyvale, CA 90292, USA
{parundek,knoblock,ambite}@usc.edu

Abstract. Despite the recent growth in the size of the Linked Data Cloud, the absence of links between the vocabularies of the sources has resulted in heterogenous schemas. Our previous work tried to find conceptual mapping between two sources and was successful in finding alignments, such as equivalence and subset relations, using the instances that are linked as equal. By using existential concepts and their intersections to define specialized classes (*restriction classes*), we were able to find alignments where previously existing concepts in one source did not have corresponding equivalent concepts in the other source. Upon inspection, we found that though we were able to find a good number of alignments, we were unable to completely cover one source with the other. In many cases we observed that even though a larger class could be defined completely by the multiple smaller classes that it subsumed, we were unable to find these alignments because our definition of *restriction classes* did not contain the disjunction operator to define a union of concepts. In this paper we propose a method that discovers alignments such as these, where a (larger) concept of the first source is aligned to the union of the subsumed (smaller) concepts from the other source. We apply this new algorithm to the Geospatial, Biological Classification, and Genetics domains and show that this approach is able to discover numerous concept coverings, where (in most cases) the subsumed classes are disjoint. The resulting alignments are useful for determining the mappings between ontologies, refining existing ontologies, and finding inconsistencies that may indicate that some instances have been erroneously aligned.

1 Introduction

The Web of Linked Data has seen huge growth in the past few years. As of September 2011, the Linked Open Data Cloud has grown to a size of 31.6 billion triples. This includes a wide range of data sources belonging to the government (42 %), geographic (19.4 %), life sciences (9.6 %) and other domains[1]. A common way that the instances in these sources are linked to others is the use of the *owl:sameAs* property. Though the size of Linked Data Cloud seems to be

[1] http://www4.wiwiss.fu-berlin.de/lodcloud/state/.

E. Simperl et al. (Eds.): ESWC 2012 Satellite Events, LNCS 7540, pp. 3–15, 2015.
DOI: 10.1007/978-3-662-46641-4_1

increasing drastically (10 % over the 28.5 billion triples in 2010), inspection of the sources at the ontology level reveals that only a few of them (15 out of the 190 sources) have some mapping of the vocabularies. For the success of the Semantic Web, it is important that these heterogenous schemas be linked. As described in our previous papers on Linking and Building Ontologies of Linked Data [8] and Aligning Ontologies of Geospatial Linked Data [7], an extensional technique can be used to generate alignments between the ontologies behind these sources. In these previous papers, we introduced the concept of *restriction classes*, which is the set of instances that satisfy a *conjunction* of value restrictions on properties (*property-value pairs*).

Though our algorithm was able to identify a good number of alignments, it was unable to completely cover one source with the classes in the other source. Upon closer look, we found that most of these alignments that we missed did not have a corresponding *restriction class* in the other source, and instead subsumed multiple *restriction classes*. While reviewing these subset relations, we discovered that in many cases the union of the smaller classes completely covered the larger class. In this paper, we describe how we extend our previous work to discover such concept coverings by introducing more expressive set of class descriptions (unions of value restrictions)[2]. In most of these coverings, the smaller classes are also found to be disjoint. In addition, further analysis of the alignments of these coverings provides a powerful tool to discover incorrect links in the Web of Linked Data, which can potentially be used to point out and rectify the inconsistencies in the instance alignments.

This paper is organized as follows. First, we describe the Linked Open Data sources that we try to align in the paper. Second, we briefly review our alignment algorithm from [8] along with the limitations of the results that were generated. Third, we describe our approach to finding alignments between unions of restrictions classes. Fourth, we describe how outliers in these alignments help to identify inconsistencies and erroneous links. Fifth, we describe the experimental results on union alignments over additional domains. Finally, we compare against related work, and discuss our contributions and future work.

2 Sources Used for Alignments

Linked Data, by definition, links the instances of multiple sources. Often, sources conform to different, but related, ontologies that can also be meaningfully linked [8]. In this section we describe some of these sources from different domains that we try to align, instances in which are linked using an equivalence property like *owl:sameAs*.

Linking *GeoNames* with places in *DBpedia:* *DBpedia* (dbpedia.org) is a knowledge base that covers multiple domains including around 526,000 places and other geographical features from the Geospatial domain. We try to align

[2] This work is an extended version of our workshop paper [6]. We have extended the method to find coverings in the Biological Classification and Genetics domains.

the concepts in *DBpedia* with *GeoNames* (geonames.org), which is a geographic source with about 7.8 million things. It uses a flat-file like ontology, where all instances belong to a single concept of *Feature*. This makes the ontology rudimentary, with the type data (e.g. mountains, lakes, etc.) about these geographical features instead in the *Feature Class & Feature Code* properties.

Linking *LinkedGeoData* with places in *DBpedia:* We also try to find alignments between the ontologies behind *LinkedGeoData* (linkedgeodata.org) and *DBpedia*. *LinkedGeoData* is derived from the *Open Street Map* initiative with around 101,000 instances linked to *DBpedia* using the *owl:sameAs* property.

Linking species from *Geospecies* with *DBpedia:* The *Geospecies* (geospecies.org) knowledge base contains species belonging to plant, animal, and other kingdoms linked to species in *DBpedia* using the *skos:closeMatch* property. Since the instances in the taxonomies in both these sources are the same, the sources are ideal for finding the alignment between the vocabularies.

Linking genes from *GeneID* with *MGI:* The Bio2RDF (bio2rdf.org) project contains inter-linked life sciences data extracted from multiple data-sets that cover genes, chemicals, enzymes, etc. We consider two sources from the Genetics domain from Bio2RDF, *GeneID* (extracted from the National Center for Biotechnology Information database) and *MGI* (extracted from the Mouse Genome Informatics project), where the genes are marked equivalent.

Although we provide results of the above four mentioned alignments in Section 4, in the rest of this paper we explain our methodology by using the alignment of *GeoNames* with *DBpedia* as an example.

3 Aligning Ontologies on the Web of Linked Data

First, we briefly describe our previous work on finding subset and equivalent alignments between *restriction classes* from two ontologies. Then, we describe how to use the subset alignments to finding more expressive *union alignments*. Finally, we discuss how outliers in these union alignments often identify incorrect links in the Web of Linked Data.

3.1 Our Previous Work on Aligning Ontologies of Linked Data

In [8] we introduced the concept of *restriction classes* to align extensional concepts in two sources. A *restriction class* is a concept that is derived extensionally and defined by a conjunction of value restrictions for properties (called *property-value pairs*) in a source. Such a definition helps overcome the problem of aligning rudimentary ontologies with more sophisticated ones. For example, *GeoNames* only has a single concept (*Feature*) to which all of its instances belong, while *DBpedia* has a rich ontology. However, *Feature* has several properties that can be used to define more meaningful classes. For example, the set of instances in *GeoNames* with the value *PPL* in the property *featureCode*, nicely aligns with the instances of *City* in *DBpedia*.

Our algorithm explored the space of *restriction classes* from two ontologies and was able to find equivalent and subset alignments between these *restriction classes*. Fig. 1 illustrates the instance sets considered to score an alignment hypothesis. We first find the instances belonging to the *restriction class* r_1 from the first source and r_2 from the second source. We then compute the *image* of r_1 (denoted by $I(r_1)$), which is the set of instances from the second source linked to instances in r_1 (dashed lines in the figure). By comparing r_2 with the intersection of $I(r_1)$ and r_2 (shaded region), we can determine the relationship between r_1 and r_2. We defined two metrics P and R, as the ratio of $|I(r_1) \cap r_2|$ to $|I(r_1)|$ and $|r_2|$ respectively, to quantify set-containment relations. For example, two classes are equivalent if $P = R = 1$. In order to allow a certain margin of error induced by the data-set, we used the relaxed versions *P'* and *R'* as part of our scoring mechanism. In this case, two classes were considered equivalent if $P' > 0.9$ and $R' > 0.9$ For example, consider the alignment between *restriction classes* (*lgd:gnis%3AST_alpha*=NJ) from *LinkedGeoData* and (*dbpedia:Place#type*=http://dbpedia.org/resource/City_(New_Jersey)) from *DBpedia*. Based on the extension sets, our algorithm finds $|I(r_1)| = 39$, $|r_2| = 40$, $|I(r_1) \cap r_2| = 39$, $R' = 0.97$ and $P' = 1.0$. Based on our error margins, we assert the alignment as equivalent in an extensional sense. The exploration of the space of alignments and the scoring procedure is described in detail in [8].

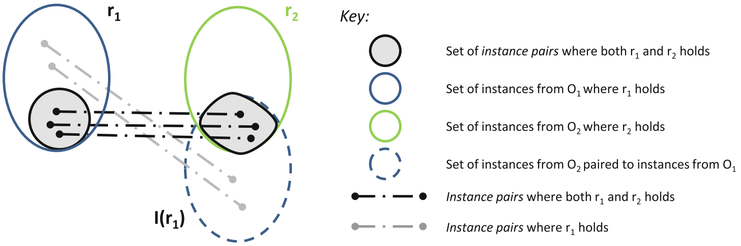
r_1
r_2
$I(r_1)$
Key:
Set of *instance pairs* where both r_1 and r_2 holds
Set of instances from O_1 where r_1 holds
Set of instances from O_2 where r_2 holds
Set of instances from O_2 paired to instances from O_1
Instance pairs where both r_1 and r_2 holds
Instance pairs where r_1 holds

Fig. 1. Comparing the linked instances from two ontologies.

Though the approach produced a large number of equivalent alignments, we were not able to find a complete coverage because some *restriction classes* did not have a corresponding equivalent *restriction class* and instead subsumed multiple smaller *restriction classes*. For example, in the *GeoNames* and *DBpedia* alignment, we found that {*rdf:type=dbpedia:EducationalInstitution*} from *DBpedia* subsumed {*geonames:featureCode=S.SCH*}, {*geonames:featureCode=S.SCHC*} and {*geonames:featureCode=S.UNIV*} (i.e. Schools, Colleges and Universities from *GeoNames*). We discovered that taken together, the union of these three *restriction classes* completely define *rdf:type=dbpedia:EducationalInstitution*. To find such previously undetected alignments we decided to extend the expressivity of our *restriction classes*by introducing a disjunction operator to detect concept coverings completely.

3.2 Identifying Union Alignments

In our current work, we use the subset and equivalent alignments generated by the previous work to try and align a larger class from one ontology with a union of smaller subsumed *restriction classes* in the other ontology. Since the problem of finding alignments with conjunctions and disjunction of *property-value pairs* of *restriction classes* is combinatorial in nature, we focus only on subset and equivalence relations that map to an *restriction classes* with a single *property-value pair*. This helps us find the simplest definitions of concepts and also makes the problem tractable. Alignments generated by our previous work that satisfy the single *property-value pair* constraint are first grouped according to the subsuming *restriction classes* and then according to the property of the smaller classes. Since *restriction classes* are constructed by forming a set of instances that have one of the properties restricted to a single value, aggregating *restriction classes* from the group according to their properties builds a more intuitive definition of the union. We can now define the disjunction operator that constructs the union concept from the smaller *restriction classes* in these sub-groups. The disjunction operator is defined for *restriction classes*, such that *(i)* the concept formed by the disjunction of the classes represents the union of their set of instances, *(ii)* each of the classes that are aggregated contain only a single *property-value pair* and *(iii)* the property for all those *property-value pairs* is the same. We then try to detect the alignment between the larger common *restriction class* and the union by using an extensional approach similar to our previous paper. We call such an alignment a hypothesis *union alignment*.

We define U_S as the set of instances that is the union of individual smaller *restriction classes* Union(r_2); U_L as the image of the larger class by itself, Img(r_1)); and U_A as the overlap between these sets, union($Img(r_1) \cap r_2$)). We check whether the larger *restriction class* is equivalent to the union concept by using scoring functions analogous to P' & R' from our previous paper. The new scoring mechanism defines P'_U as $\frac{|U_A|}{|U_S|}$ and R'_U as $\frac{|U_A|}{|U_L|}$ with relaxed scoring assumptions as in P' & R'. To accommodate errors in the data-set, we consider it a complete coverage when the score is greater than a relaxed score of 0.9. That is, the hypothesis *union alignment* is considered equivalent if $P'_U > 0.9$ & $R'_U > 0.9$. Since by construction, each of the subset already satisfies $P' > 0.9$, then we are assured that P'_U is always going to be greater than 0.9. Thus, a *union alignment* is equivalent if $R'_U > 0.9$.

Figure 2 provides an example of the approach. Our previous algorithm finds that {*geonames:featureCode = S.SCH*}, {*geonames:featureCode = S.SCHC*}, {*geonames:featureCode = S.UNIV*} are subsets of {*rdf:type=dbpedia:EducationalInstitution*}. As can be seen in the Venn diagram in Fig. 2, U_L is *Img*({*rdf:type = dbpedia:EducationalInstitution*}*)*, U_S is {*geonames:featureCode = S.SCH*} ∪ {*geonames:featureCode = S.SCHC*} ∪ {*geonames:featureCode = S.UNIV*}, and U_A is the intersection of the two. With the educational institutions example, R'_U for the alignment of *dbpedia:EducationalInstitution* to the union of *S.SCH, S.SCHC & S.UNIV* is 0.98. We can thus confirm the hypothesis and consider this *union alignment* equivalent. Section 4 shows additional examples of *union alignments*.

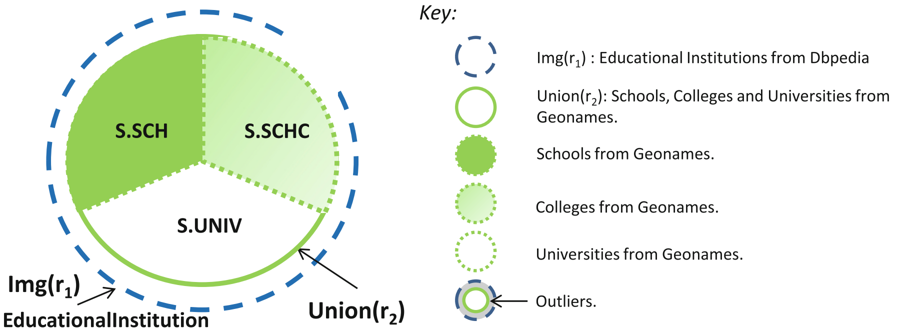
S.SCH
S.SCHC
S.UNIV
Img(r_1)
EducationalInstitution
Union(r_2)
Key:
Img(r_1) : Educational Institutions from Dbpedia
Union(r_2): Schools, Colleges and Universities from Geonames.
Schools from Geonames.
Colleges from Geonames.
Universities from Geonames.
Outliers.

Fig. 2. Spatial covering of Educational Institutions from *DBpedia*

3.3 Using Outliers in Union Mappings to Identify Linked Data Errors

The computation of *union alignments* allows for a margin of error in the subset computation. It turns out that the outliers, the instances that do not satisfy the *restriction classes* in the alignments, are often due to incorrect links. Thus, our algorithm also provides a novel method to curate the Web of Linked Data.

Consider the outlier found in the {*dbpedia:country* = *Spain*} ≡ {*geonames:countryCode* = *ES*} alignment. Of the 3918 instances of *dbpedia:country=Spain*, 3917 have a link to a *geonames:countryCode=ES*. The one instance not having country code ES has an assertion of country code IT (Italy) in *GeoNames*. The algorithm would flag this situation as a possible linking error, since there is overwhelming support for the ES being the country code of Spain. A more interesting case occurs in the alignment of {*rdf:type* = *dbpedia:EducationalInstitution*} to {*geonames:featureCode* ∈ {*S.SCH, S.SCHC, S.UNIV*}. For {*rdf:type* = *dbpedia:EducationalInstitution*}, 396 instances out of the 404 Educational Institutions were accounted for as having their *geonames:featureCode* as one of *S.SCH, S.SCHC or S.UNIV*. From the 8 outliers, 1 does not have a *geonames:featureCode* property asserted. The other 7 have their feature codes as either S.BLDG (3 buildings), S.EST (1 establishment), S.HSP (1 hospital), S.LIBR (1 library) or S.MUS (1 museum). This case requires more sophisticated curation and the outliers may indicate a case for multiple inheritance. For example, the hospital instance in geonames may be a medical college that could be classified as a university.

Our union alignment algorithm is able to detect similar other outliers and provides a powerful tool to quickly focus on links that require human curation, or that could be automatically flagged as problematic, and provides evidence for the error.

4 Experimental Results

The results of union alignment algorithm over the four pairs of sources we consider appear in Table 1. In total, the 7069 union alignments explained (covered) 77966 subset alignments, for a compression ratio of 90 %.

Table 1. Union alignments found in the 4 source pairs

Source1	Source2	Union alignments 12 (Subset Alignments 21)	Union alignments 21 (Subset Alignments 21)	Total union alignments
GeoNames	*DBpedia*	434 (2197)	318 (7942)	752
LinkedGeoData	*DBpedia*	2746 (12572)	3097 (48345)	5843
Geospecies	*DBpedia*	191 (1226)	255 (2569)	446
GeneID	*MGI*	6 (29)	22 (3086)	28

The resulting alignments were intuitive. Some interesting examples appear in Tables 2, 3 and 4. In the tables, for each *union alignment*, column 2 describes the large *restriction class* from $ontology_1$ and column 3 describes the union of the (smaller) classes on $ontology_2$ with the corresponding property and value set. The score of the union is noted in column 4 ($R'_U = \frac{|U_A|}{|U_L|}$) followed by $|U_A|$ and $|U_L|$ in columns 5 and 6. Column 7 describes the outliers, i.e. values of v_2 that form *restriction classes* that are not direct subsets of the larger *restriction class*. Each of these outliers also has a fraction with the number of instances that belong to the intersection upon the the number of instances of the smaller *restriction class* (or $\frac{|Img(r_1) \cap r_2|}{|r_2|}$). It can be seen that the fraction is less than our relaxed subset score. If the value of this fraction was greater than the relaxed subset score (i.e. 0.9), the set would have been included in column 3 instead. The last column mentions how many of the total U_L instances were we able to explain using U_A and the outliers. For example, the *union alignment* #1 of Table 2 is the Educational Institution example described before. It shows how educational institutions from *DBpedia* can be explained by schools, colleges and universities in *GeoNames*. Column 4, 5 and 6 explain the alignment score R'_U (0.98), the size U_A (396) and the size of U_L (404). Outliers (S.BLDG, S.EST, S.LIBR, S.MUS, S.HSP) along with their P' fractions appear in column 7. We were able to explain 403 of the total 404 instances (see column 8).

We find other interesting alignments, a representative few of which are shown in the tables. In some cases, the *union alignments* found were intuitive because of an underlying hierarchical nature of the concepts involved, especially in case of alignments of administrative divisions in geospatial sources and alignments in the biological classification taxonomy. For example, #3 highlights alignments that reflect the containment properties of administrative divisions. Other interesting types of alignment were also found. For example #7 tries to map two non-similar concepts. It explains the license plate codes found in the state (bundesland) of Saarland[3]. Due to lack of space, we explain the other *union alignments* alongside in the tables. The complete set of alignments discovered by our algorithm are available on our group page.[4]

[3] http://www.europlates.com/publish/euro-plate-info/german-city-codes.

[4] http://www.isi.edu/integration/data/UnionAlignments.

Table 2. Example alignments from the *GeoNames-DBpedia LinkedGeoData-DBpedia*.

#	$\{r_1\}$	$p_2 \in \{v_2\}$	$R'_U = \frac{\lvert U_A\rvert}{\lvert U_L\rvert}$	$\lvert U_A\rvert$	$\lvert U_L\rvert$	***Outliers***	***# Explained Instances***
***DBpedia* (larger) - *GeoNames* (smaller)**							
1	{*rdf:type* = *dbpedia:EducationalInstitution*}	*geonames:featureCode* ∈ {S.SCH, S.SCHC, S.UNIV}	0.9801	396	404	S.BLDG (3/122), S.EST (1/13), S.LIBR (1/7), S.HSP (1/31), S.MUS (1/43)	403
	As described in Section 4, Schools, Colleges and Universities in *GeoNames*make Educational Institutions in *DBpedia*						
2	{*dbpedia:country* = *dbpedia:Spain*}	*geonames:countryCode* = *ES*	0.9997	3917	3918	IT (1/7635)	3918
	The concepts for the country Spain are equal in both sources. The only outlier has it's country as Italy, an erroneous assertion.						
3	*dbpedia:region* = *dbpedia:Basse-Normandie*	*geonames:parentADM2* ∈ {geonames:2989247, geonames:2996268, geonames:3029094}	1.0	754	754		754
	We confirm the hierarchical nature of administrative divisions with alignments between administrative units at two different levels.						
4	{*rdf:type* = *dbpedia:Airport*}	*geonames:featureCode* ∈ {S.AIRB, S.AIRP}	0.9924	1981	1996	S.AIRF (9/22), S.FRMT (1/5), S.SCH (1/404), S.STNB (2/5) S.STNM (1/36), T.HLL (1/61)	1996
	In alignmening airports, an airfield should have been an an airport. However, there was not enough instance support.						
***GeoNames* (larger) - *DBpedia* (smaller)**							
5	{*geonames:countryCode* = *NL*}	*dbpedia:country* ∈ {dbpedia:The_Netherlands, dbpedia:Flag_of_the _Netherlands.svg, dbpedia:Netherlands}	0.9802	1939	1978	dbpedia:Kingdom_of _the_Netherlands	1940
	The Alignment for Netherlands should have been as straightforward as #2. However we have possible alias names, such as *The Netherlands* and *Kingdom of Netherlands*, as well a possible linkage error to *Flag of the Netherlands.svg*						
6	{*geonames:countryCode* = *JO*}	*dbpedia:country* ∈ {dbpedia:Jordan, dbpedia:Flag_of_Jordan.svg}	0.95	19	20		20
	The error pattern in #5 seems to repeat systematically, as can be seen from this alignment for the coutry of Jordan.						
***DBpedia* (larger) - *LinkedGeoData* (smaller)**							
7	{*dbpedia:bundesland* = *Saarland*}	*lgd:OpenGeoDBLicensePlate-Number* ∈ { HOM, IGB, MZG, NK, SB, SLS, VK, WND}	0.93	46	49		46
	Our algorithm also produces interesting alignments between different properties. In this case, we find 8 of the 10 license plates in the state of Saarland						

Table 3. Example alignments from the *LinkedGeoData-DBpedia*, *Geospecies-DBpedia*

#	$\{r_1\}$	$p_2 \in \{v_2\}$	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
8	{*rdf:type*, *dbpedia:EducationalInstitution*}	*rdf:type* ∈ {lgd:Amenity, lgd:K2543, lgd:School, lgd:University, lgd:WaterTower}	0.9901	2609	2610		2609
	Educational Institutions in *DBpedia* can be explained with classes in *LinkedGeoData*. An example of an incorrent alignment, a water tower has been linked to as an educational institution.						
LinkedGeoData **(larger) -** ***DBpedia*** **(smaller)**							
9	{*lgd:gnisST_alpha = NJ*}	*dbpedia:subdivisionName* ∈ {Atlantic, Burlington, {Cape May, Hudson, Hunterdon, Monmoth, New Jersey, Ocean, Passaic}	1.0	214	214		214
	Due to missing instance alignments, this *union alignment* incorrectly claims that the state of New Jersey is composed of 9 counties while actually it has 21.						
10	{*rdf:type = lgd:Waterway*}	*rdf:type* ∈ dbpedia:River dbpedia:Stream}	0.97	33	34	dbpedia:Place(1/94989)	34
	Waterways in *LinkedGeoData* as equal to the union of streams and rivers from *DBpedia*						
DBpedia **(larger) -** ***Geospecies*** **(smaller)**							
11	{*rdf:type = dbpedia:Amphibian*} *dbpedia:Amphibian* }	*geospecies:hasOrderName* ∈ {Anura, Caudata, Gymnophionia}	0.99	90	91	Testudines (1/7)	91
	Species from *Geospecies* with the order names Anura, Caudata & Gymnophionia are all Amphibians We also find inconsistancies due to misaligned instances, e.g. one Turtle (Testidune) was classified as amphibian.						
12	{*rdf:type = dbpedia:Salamander*}	{*geospecies:hasOrderName* = Caudata}	0.94	16	17	Testudines (1/7)	17
	Upon further inspection of #11, we find that the culprit is a Salamander						
Geospecies **(larger) -** ***DBpedia*** **(smaller)**							
13	{*rdf:type = dbpedia:Plant*}	{*geospecies:inKingdom* = geospecies:kingdoms/Ab}	0.99	1874	1876	geospecies:kingdoms/Ac(1/8)	1875
	The Kingdom Plantae, from both sources, almost matches perfectly. The only inconsistant instance happens to be a fungus.						

Table 4. Example alignments from the *GeneID-MGI*

#	{r_1}	$p_2 \in \{v_2\}$	$R'_U = \frac{\lvert U_A\rvert}{\lvert U_L\rvert}$	$\lvert U_A\rvert$	$\lvert U_L\rvert$	***Outliers***	***# Explained Instances***
14	{*geospecies:inOrder = geospecies:orders/jtSaY*}	*dbpedia:ordo* ∈ {dbpedia:Carnivora, dbpedia:Carnivore}	0.99	247	247		247
	Inconsistancies in the object values can also be seen - Carnivores from *Geospecies* are aligned with both : Carnivora & Carnivore.						
15	{*geospecies:hasOrderName = Chiroptera*}	*dbpedia:ordo* ∈ {Chiroptera@en, dbpedia:Bat}	1	111	111		111
	We can detect that species with order Chiroptera correctly belong to the order of Bats. Unfortunatey, due to values of the property being the literal "Chiropta@en", the alignment is not clean.						
***GeneID* (larger) - *MGI* (smaller)**							
16	{*bio2rdf:subType = pseudo*}	{*bio2rdf:subType* = Pseudogene}	0.93	5919	6317	Gene (318/24692)	6237
	Due to the absence of a clear hierarchy, we found only a few hierarchical relations. For example, alignments of the classes Pseudogenes.						
17	{*bio2rdf:xTaxon = taxon:10090*}	*bio2rdf:subType* ∈ {Complex Cluster/Region, DNA Segment, Gene, Pseudogene}	1	30993	30993		30993
	The Mus Musculus (house mouse) taxonomy is completely composed of complex clusters, DNA segments, Genes and Pseudogenes .						
***MGI* (larger) - *GeneID* (smaller)**							
18	{*bio2rdf:subType = Pseudogene*}	*bio2rdf:subType = pseudo*	0.94	5919	6297	other (4/230) protein-coding (351/39999) unknown(23/570)	6297
	Inconsistancies are also evident as the values pseudo and Pseudogene are used to denote the same thing.						
19	{*mgi:genomeStart = 1*}	*geneid:location* ∈ {1, 1 0.0 cM, 1 1.0 cM, 1 10.4 cM, ...}	0.98	1697	1735	""(37/1048) 5 (1/52)	1735
20	{*mgi:genomeStart = X*}	*geneid:location* ∈ {X, X 0.5 cM, X 0.8 cM, X 1.0 cM, ...}	0.99	1748	1758	""(10/1048)	1758
	We find interesting alignments like #19 & #20 , which align the genome start position in *MGI* with the location in *GeneID* As can be seen, the values of the locations (distances in centimorgans) in *GeneID* contain genome start value as a prefix. Inconsistancies are also seen, e.g. in #19 a gene that starts with 5 is misaligned and in #20, where the value is an empty string.						

Outliers. In alignments that had inconsistencies, we identified three main reasons: **(i)** *Incorrect instance alignments* - outliers arising out of possible erroneous equivalence link between instances (e.g. #4, #8, etc.), **(ii)** *Missing instance alignments* - insufficient support for coverage due to missing links between instances or missing instances (e.g. #9, etc.), **(iii)** *Incorrect values for properties* - outliers arising out of possible erroneous assertion for property (e.g. #5, #6, etc.). In the tables, we also mention the classes that these inconsistencies belong to along with their support.

5 Related Work

Ontology alignment and schema matching have been a well explored area of research since the early days of ontologies [1,3] and received renewed interest in recent years with the rise of the Semantic Web and Linked Data. Though most work done in the Web of Linked Data is on linking instances across different sources, an increasing number of authors have looked into aligning the source ontologies in the past couple of years. Jain et al. [4] describe the BLOOMS approach which uses a central forest of concepts derived from topics in Wikipedia. An update to this is the BLOOMS+ approach [5] that aligns Linked Open Data ontologies with an upper-level ontology called Proton. BLOOMS & BLOOMS+ are unable to find alignments because of the small number of classes in *GeoNames* that have vague declarations. The advantage of our approach over these is that our use of *restriction classes* is able to find a large set of alignments in cases like aligning *GeoNames* with *DBpedia* where Proton fails due to a rudimentary ontology. Cruz et al. [2] describe a dynamic ontology mapping approach called *AgreementMaker* that uses similarity measures along with a mediator ontology to find mappings using the labels of the classes. From the subset and equivalent alignment between *GeoNames*(10 concepts) and *DBpedia*(257 concepts), AgreementMaker was able to achieve a precision of 26 % and a recall of 68 %. We believe that since their approach did not consider unions of concepts, it would not have been able to find alignments like the Educational Institutions example (#1) by using only the labels and the structure of the ontology, though a thorough comparison is not possible. In our work, we find equivalent relations between a concept on one side and a union of concepts on another side. *CS*R [9] is a similar work to ours that tries to align a concept from one ontology to a union of concepts from the other ontology. In their approach, the authors describe how the similarity of properties are used as features in predicting the subsumption relationships. It differs from our approach in that it uses a statistical machine learning approach for detection of subsets rather than the extensional approach. An approach that uses statistical methods for finding alignments, similar to our work, has also been described in Völker et al. [10]. This work induces schemas for RDF data sources by generating OWL2 axioms using an intermediate associativity table of instances and concepts (called *transaction data-sets*) and mining associativity rules from it.

6 Conclusions and Future Work

We described an approach to identifying *union alignments* in data sources on the Web of Linked Data from the Geospatial, Biological Classification and Genetics domains. By extending our definition of *restriction classes* with the disjunction operator, we are able to find alignments of union concepts from one source to larger concepts from the other source. Our approach produce coverings where concepts at different levels in the ontologies of two sources can be mapped even when there is no direct equivalence. We are also able to find outliers that enable us to identify inconsistencies in the instances that are linked by looking at the alignment pattern. The results provide deeper insight into the nature of the alignments of Linked Data.

As part of our future work we want to try to find a more complete descriptions for the sources. Our preliminary findings show that the results of this paper can be used to find patterns in the properties. For example, the *countryCode* property in *GeoNames* is closely associated with the *country* property in *DBpedia*, though their ranges are not exactly equal. We believe that an in-depth analysis of the alignment of ontologies of sources is warranted with the recent rise in the links in the Linked Data cloud. This is an extremely important step for the grand Semantic Web vision.

References

1. Bernstein, P., Madhavan, J., Rahm, E.: Generic schema matching, ten years later. Proc. VLDB Endow. **4**(11), 695–701 (2011)
2. Cruz, I., Palmonari, M., Caimi, F., Stroe, C.: Towards on the go matching of linked open data ontologies. In: Workshop on Discovering Meaning On The Go in Large Heterogeneous Data, p. 37 (2011)
3. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
4. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology alignment for linked open data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 402–417. Springer, Heidelberg (2010)
5. Jain, P., Yeh, P.Z., Verma, K., Vasquez, R.G., Damova, M., Hitzler, P., Sheth, A.P.: Contextual ontology alignment of LOD with an upper ontology: a case study with proton. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 80–92. Springer, Heidelberg (2011)
6. Parundekar, R., Ambite, J.L., Knoblock, C.A.: Aligning unions of concepts in ontologies of geospatial linked data. In: Proceedings of the Terra Cognita 2011 Workshop in Conjunction with the 10th International Semantic Web Conference, Bonn, Germany (2011)
7. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Aligning geospatial ontologies on the linked data web. In: Proceedings of the GIScience Workshop on Linked Spatiotemporal Data, Zurich, Switzerland (2010)
8. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Linking and building ontologies of linked data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 598–614. Springer, Heidelberg (2010)

9. Spiliopoulos, V., Valarakos, A.G., Vouros, G.A.: *CSR*: Discovering subsumption relations for the alignment of ontologies. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 418–431. Springer, Heidelberg (2008)
10. Völker, J., Niepert, M.: Statistical schema induction. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 124–138. Springer, Heidelberg (2011)