Discovering Concept Coverings in Ontologies of Linked Data Sources

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, CA 90292
{parundek,knoblock,ambite}@usc.edu

Abstract. Despite the increase in the number of linked instances in the Linked Data Cloud in recent times, the absence of links at the concept level has resulted in heterogenous schemas, challenging the interoperability goal of the Semantic Web. In this paper, we address this problem by finding alignments between concepts from multiple Linked Data sources. Instead of only considering the existing concepts present in each ontology, we hypothesize new composite concepts defined as disjunctions of conjunctions of (RDF) types and value restrictions, which we call *restriction classes*, and generate alignments between these composite concepts. This extended concept language enables us to find more complete definitions and to even align sources that have rudimentary ontologies, such as those that are simple renderings of relational databases. Our concept alignment approach is based on analyzing the extensions of these concepts and their linked instances. Having explored the alignment of conjunctive concepts in our previous work, in this paper, we focus on concept coverings (disjunctions of *restriction classes*). We present an evaluation of this new algorithm to Geospatial, Biological Classification, and Genetics domains. The resulting alignments are useful for refining existing ontologies and determining the alignments between concepts in the ontologies, thus increasing the interoperability in the Linked Open Data Cloud.

1 Introduction

The Web of Linked Data has grown significantly in the past few years – 31.6 billion triples as of September 2011. This includes a wide range of data sources from 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 through the *owl:sameAs* property. Though the size of Linked Data Cloud is increasing steadily (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) include mappings between their ontologies. Since interoperability is crucial to the success of the Semantic Web, it is essential that these heterogenous schemas, the result of a de-centralized approach to the generation of data

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

P. Cudré-Mauroux et al. (Eds.): ISWC 2012, Part I, LNCS 7649, pp. 427–443, 2012.

and ontologies, also be linked. The problem of schema linking, such as schema matching in databases and ontology alignment in the Semantic Web, has been well researched [5,1,4]. As in Instance-based Matching [4,3,7], we follow an extensional approach to generating the alignments. The novelty of our approach consists of generating new concept hypotheses beyond the concepts originally present in the ontologies, and aligning these extended concepts by exploiting the linked instances in the Linked Data Cloud.

The problem of finding alignments in ontologies of Linked Data sources is non-trivial, since there might not be one-to-one concept equivalences. In some sources the ontology is extremely rudimentary, for example *GeoNames* has only one class - *geonames:Feature*, and the alignment of such an ontology with a well defined one, such as *DBpedia*, is not particularly useful. In order to be successful in linking ontologies, we need to generate more expressive concepts. The necessary information to do this is often present in the properties and values of the instances in the sources. For example, in *GeoNames* the values of the *featureCode* and *featureClass* properties provide useful concept constructors, which can be aligned with existing concepts in *DBpedia*, so that we have that the concept *geonames:featureCode=P.PPL* (populated place) aligns to *dbpedia:City*. Therefore, our approach explores the space of concepts defined by value restrictions, which we will call *restriction classes* in the reminder of the paper. A *value restriction* is a concept constructor present in expressive description logics, such as OWL-DL ($\mathcal{SHOIN}(\mathcal{D})$) [6]. We consider class assertions (rdf:type) and value restrictions on both object and data properties, which we will represent uniformly as $\{p = v\}$, where either p is an object property and v is a resource (including rdf:type=Class), or p is a data property and v is a literal. We consider two *restriction classes* equal if their respective instance sets can be identified as equal after following the *owl:sameAs* links.

In our previous work [10], we explored conjunctive *restriction classes*. In this paper, we explore disjunctive *restriction classes*. Specifically, we focus on concept coverings where a larger concept from one source can be explained by (i.e., is extensionally equivalent to) the union of multiple smaller classes in the other source. Our approach finds alignments based on the extensions of the concepts, that is, the sets of instances satisfying the definitions of the restriction classes. We believe that this is an important feature of our approach in that it allows one to understand the relationships in the *actual* linked data and their corresponding ontologies. The alignments generated can readily be used for modeling and understanding the sources since we are modeling what the sources actually contain as opposed as to what an ontology disassociated from the data appears to contain based on the class name or description.

This paper is organized as follows. First, we describe the Linked Open Data sources that we align. Second, we present the alignment algorithm that consists of two steps: finding initial equivalence and subset relations, and then discovering *concept coverings* using disjunctions of *restriction classes*. Third, we describe representative alignments discovered by our approach and present an evaluation of the results. An interesting outcome of our algorithm is that it identifies

inconsistencies and possible errors in the linked data, and provides a method for automatically curating the Linked Data Cloud. Finally, we compare against related work, and discuss our contributions and future work.[2]

2 Sources Used for Alignments

In the Linked Open Data Cloud, sources often conform to different, but related, ontologies that can also be meaningfully linked [2,9,10]. In this section we describe some of these sources from different domains that we 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 align concepts in *DBpedia* with *GeoNames* (geonames.org), which is a geographic source with about 7.8 million geographical features. GeoNames uses a rudimentary flat-file like ontology, where all instances belong to a single concept of *Feature*, with the type data (e.g. mountains, lakes, etc.) encoded in the *featureClass and featureCode* properties.

Linking *LinkedGeoData* with Places in *DBpedia*: We also 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* knowledge base (lod.geospecies.org) contains a taxonomic classification of living organisms linked to species in *DBpedia* using the *skos:closeMatch* property. Since these sources have many species in common, they are ideal for finding alignments 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.

In Section 4 we provide results for the four alignment experiments described above. In the rest of this paper we explain our methodology, which is source independent, by using the alignment of *GeoNames* with *DBpedia* as an example.

3 Finding Concept Coverings across Ontologies

We use a two step approach to find *concept coverings*. First, we extract atomic equivalent and subset alignments from the two sources where the *restriction*

[2] This paper is an extended version of our workshop paper [11]. We have added more formal descriptions, explanations of the algorithms and detailed evaluation.

class on each side contains a single *property-value pair* and no conjunction or disjunctions. These are the simplest alignments that can be defined. We then use these to find concept coverings by describing a larger concept with a union of smaller ones using set containment.

We also discuss how our alignment approach detects outliers, which often indicate missing or incorrect links, and provides a powerful tool to curate the Linked Data cloud.

3.1 Finding Alignments with Atomic *Restriction Classes*

As a precursor to finding *concept coverings* between the two sources, our algorithm first finds alignments where the *restriction classes* on each side of the alignment are atomic - i.e. have one *property-value pair* each. In our previous work [10], we used a similar approach to find alignments between the ontologies where a conjunction of *restriction classes* was aligned with its equivalent concept in the other source. In this paper we focus on atomic *restriction classes*. Since we do not need to find alignments of conjunctive *restriction classes*, the search space is polynomial rather than combinatorial.

The sources are first prepared for exploration by performing an inner join on the equivalence property (e.g., *owl:sameAs*) and optimized by removing inverse-functional properties. Then, the following algorithm is used to find the alignments between atomic *restriction classes*.

for all p_1 in $Source_1$ and distinct v_1 associated with p_1 **do**
  $r_1 \leftarrow$ *restriction class* $\{p_1 = v_1\}$ containing all instances where $p_1 = v_1$
  $Img(r_1) \leftarrow$ Find all corresponding instances from $Source_2$ to those in r_1, linked by *owl:sameAs*
  for all p_2 in $Source_2$ and distinct v_2 associated with p_2 **do**
    $r_2 \leftarrow$ *restriction class* $\{p_2 = v_2\}$ containing all instances where $p_2 = v_2$
    $P \leftarrow \frac{|Img(r_1) \cap r_2|}{|r_2|}$, $R \leftarrow \frac{|Img(r_1) \cap r_2|}{|r_1|}$
    if $P \geq \theta$ **then** $alignment(r_1, r_2) \leftarrow r_1 \subset r_2$
    end if
    if $R \geq \theta$ **then** $alignment(r_1, r_2) \leftarrow r_2 \subset r_1$
    end if
    if $P \geq \theta$ *and* $R \geq \theta$ **then** $alignment(r_1, r_2) \leftarrow r_1 \equiv r_2$
    end if
  end for
end for

Fig. 1 illustrates the set comparison operations of our algorithm. In order to allow a certain margin of error induced by the data set, we use $P \geq \theta$ and $R \geq \theta$ (instead of $P = 1$ and $R = 1$, which would hold if there were no error or missing links) in our score function. In our experiments we used a threshold $\theta = 0.9$, which was determined empirically, but can be changed as desired. For example, consider the alignment between *restriction classes* {*geonames:countryCode*=ES}

from *GeoNames* and {*dbpedia:country* = *dbpedia:Spain*} from *DBpedia*. Based on the extension sets, our algorithm finds $|Img(r_1)| = 3198$, $|r_2| = 4143$, $|Img(r_1) \cap r_2| = 3917$, $R' = 0.9997$ and $P' = 0.9454$. Thus, the algorithm considers the alignment as equivalent in an extensional sense. Some alignments that do not qualify as equivalent, but with the smaller concept contained in the larger concept, qualify as subset relations. For example, we find that each of {*geonames:featureCode* = *S.SCH*}, {*geonames:featureCode* = *S.SCHC*} and {*geonames:featureCode* = *S.UNIV*} (i.e. Schools, Colleges and Universities from *GeoNames*) are subsets of {*dbpedia:EducationalInstitution*}.

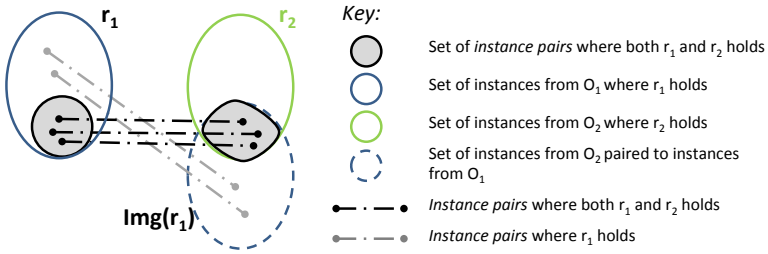

Fig. 1. Comparing the linked instances from two ontologies

Similar to our previous work [10], we also use certain optimization strategies for faster computation. For example, if we explore the properties lexicographically, the search space is reduced to half because of symmetry. To qualify as a concept, the intersection of the *restriction classes* needs to have a minimum support, which we set experimentally to ten instances.

3.2 Identifying *Concept Coverings*

In step two, we use the subclasses and equivalent alignments generated by the previous step to try and align a larger concept from one ontology with a union of smaller subsumed concepts in the other ontology. To define a larger concept, we group its subclasses from the other source that have a common property and check whether they are able to cover the larger concept. By keeping the larger *restriction class* atomic and by grouping the smaller *restriction classes* with a common property, we are able to find intuitive definitions while keeping the problem tractable. The disjunction operator that groups the smaller *restriction classes* is defined such that *i)* the concept formed by the disjunction of the classes represents the union of their set of instances, *ii)* the property for all the *property-value pairs* of the smaller aggregated classes is the same. We then try to detect the alignment between the larger concept and the union *restriction class* by using an extensional approach similar to the previous step. The algorithm for generating the hypotheses and the alignments is as follows:

for all alignments found in the previous step, with larger concepts from one source with multiple subclasses from the other source **do**
  $U_L \leftarrow$ larger *restriction class*$\{p_L = v_L\}$, and corresponding instances.
  for all smaller concepts grouped by a common property (p_S) **do**
    $U_S \leftarrow$ the union *restriction class*, and the corresponding instances of all the smaller *restriction classes*$\{p_S = \{v_1, v_2, ...\}\}$
    $U_A \leftarrow Img(U_L) \cap U_S$, $P_U \leftarrow \frac{|U_A|}{|U_S|}$, $R_U \leftarrow \frac{|U_A|}{|U_L|}$
    if $R_U \geq \theta$ **then** $alignment(r_1, r_2) \leftarrow U_L \equiv U_S$
    end if
  end for
end for

Since all smaller classes are subsets of the larger *restriction class*, $P_U \geq \theta$ by construction. We used $\theta = 0.9$ in our experiments. The smaller *restriction classes* that were omitted in the first step because of insufficient support size of their intersections, were included in constructing U_S for completeness.

Figure 2 provides an example of the approach. The first step is able to detect that alignments such as $\{$*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 Figure 2, U_L is *Img*($\{$*rdf:type* $=$ *dbpedia:EducationalInstitution*$\}$), U_S is $\{$*geonames:featureCode* $=$ *S.SCH*$\} \cup \{$*geonames:featureCode* $=$ *S.SCHC*$\} \cup \{$*geonames:featureCode* $=$ *S.UNIV*$\}$, and U_A is the intersection of the two. Upon calculation we find that R'_U for the alignment of *dbpedia:EducationalInstitution* to $\{$*geonames:featureCode*$=$ {S.SCH, S.SCHC, S.UNIV}$\}$ is 0.98. We can thus confirm the hypothesis and consider U_L & U_S equivalent. Section 4 describes these calculations and additional examples of *concept coverings*.

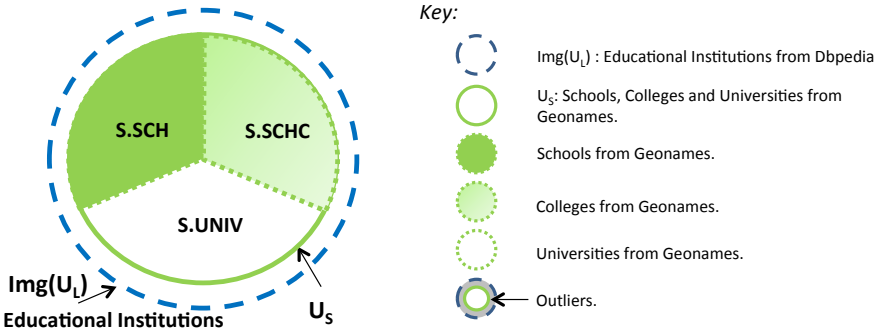

Fig. 2. Concept covering of Educational Institutions from *DBpedia*

3.3 Curating the Linked Data Cloud

It turns out that the outliers, the instances of the *restriction classes* that do not satisfy subset relations despite the error margins, are often due to incorrect and missing links or assertions. We are able to detect these, thus providing a novel method to curate the Web of Linked Data.

In the alignment of {*rdf:type = dbpedia:EducationalInstitution*} to {*geonames: featureCode* = {S.SCH, S.SCHC, S.UNIV}} we find 8 outliers (Table 6, row 1). 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.

In the {*dbpedia:country = Spain*} $\equiv$ {*geonames:countryCode* = {ES}} alignment (Table 6, row 2), one outlier instance was identified as having the country code IT (Italy) in *GeoNames*, suggesting an incorrect link/assertion. The algorithm was able to flag this situation as a possible error, since there is overwhelming support for 'ES' being the country code of Spain. 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 our *concept covering* algorithm over the four pairs of sources we consider appear in Table 1. The first step of our algorithm was able to generate about 180k equivalence and subset alignments. After running the covering algorithm, 77966 subset alignments were explained by 7069 coverings, for a compression ratio of about 11:1.

Table 1. Concept Coverings Found in the 4 Source Pairs

$Source_1$	$Source_2$	O_1-O_2: Coverings (Subset Alignments)	O_2-O_1 Coverings (Subset Alignments)	Total Coverings
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

4.1 Representative Examples of the Concept Coverings Found

Some representative examples of the *concept coverings* found are shown in Tables 6, 7 and 8. In the tables, for each *concept covering*, 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 v_2 of property p_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 over the the number of instances of the smaller *restriction class* (or $\frac{|Img(r_1) \cap r_2|}{|r_2|}$). One can see 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. $\theta = 0.9$), 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 *concept covering* #1 of Table 6 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. Thus, 403 of the total 404 instances were identified as either part of the covering or the outliers (see column 8). The remaining instance did not have a *geonames:featureCode* property asserted.

In some of the *concept coverings* discovered, the 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. For space, we explain the other *concept coverings* inside Tables 6, 7 and 8. The complete set of alignments discovered by our algorithm is available online.[3]

Outliers. In alignments, we also found inconsistencies, identified by three main reasons: **(i)** *Incorrect instance alignments* - outliers arising out of possible erroneous equivalence link between instances (e.g., in #4, a hill is linked to an airport, etc.), **(ii)** *Incorrect values for properties* - outliers arising out of possible erroneous assertion for a property (e.g. #5, #6, Flags of countries appear as values for the *country* property). In the tables, we also mention the classes that these inconsistencies belong to along with their support. We are unable to detect correct alignments if there is insufficient support for coverage due to missing links between instances or missing instances (e.g. in #9 we find a complete coverage with all instances, but it is incomplete – the state of New Jersey has 21 counties).

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

4.2 Evaluation

We present an evaluation of a random set of 642 discovered alignments across the tested source pairs to describe the precision of our approach. We checked the correctness of each of the 642 alignments manually, after verifying the completeness of *concept coverings* on websites with the relevant information. Precision was calculated as the ratio of alignments marked correct to the size of the random set. Establishing recall is difficult as finding the ground truth of all possible *concept coverings* is infeasible due to the large size of the sources and the combinatorial nature of the disjuntions. We do, however, provide an evaluation of the country alignments found in terms of precision and recall as an example.

Linking *GeoNames* with places in *DBpedia*: As shown in Table 2, out of the 752 (i.e. 434 + 328) alignments found between *GeoNames* and *DBpedia*, we evaluated 236 (i.e. 185 + 51) alignments. 152 (i.e. 127 + 25) of them were found to be correct after resolving redirects (synonyms in *DBpedia*), giving a precision of 64.40%, while 84 alignments were found to be incorrect. These 84 alignments were found to suffer common patterns of error. There are 40 alignments that had incorrect assertions of their properties. For example, in many instances in *DBpedia*, the `county` property assertion was misspelled as `country` (especially for places in UK & Ukraine), or the ".svg" file of the flag of a country appeared *dbpedia:country* value. The corresponding alignments, which we counted as incorrect, could have been properly detected if the data was cleaner. We detected only partial alignments for 14 others, where the smaller concepts left out were incorrectly classified as outliers due to insufficient support ($R < 0.9$). There were 7 partial alignments that were incorrectly detected as complete ($R > 0.9$), similar to the New Jersey example mentioned earlier. Another 14 alignments suffered from a mismatch to a parent, because of insufficient links/instances. The remaining 9 alignments had an assortment of problems in the values of properties. For example, regions inside a country (Andean Region of Colombia) appeared as value for the country property (Colombia).

Precision, recall and f-measure of Country Alignments: Since manually establishing ground truth for all possible concept coverings in the four sources is infeasible, we decided to find the precision and recall of only the country alignments we found, as an illustration. These are alignments having a common pattern, aligning a *restriction class* with a *dbpedia:country* property with other *restriction classes* featuring *geonames:countryCode* property or vice-versa. A ground truth was established by manually checking what possible country alignments were present in the two sources. Even then, establishing the ground truth needed some insight. For example, Scotland, England, Wales, Northern Ireland & the United Kingdom are all marked as countries in *DBpedia*, while in *GeoNames*, the only corresponding country is the United Kingdom. In cases like these, we decided to relax the evaluation constraint of having and alignment with a country from either of these, as correct. Another similar difficulty was in cases where militarily occupied territories were marked as countries (e.g. Golan Heights occupied by Israel is marked as *dbpedia:country*).

Table 2. Linking *GeoNames* with places in *DBpedia*

Description of Pattern Observed	Alignments w/ larger class from *GeoNames*	Alignments w/ larger class from *DBpedia*
Total # Alignments	434	328
# Alignments Evaluated	**185**	**51**
Correct	**127**	**25**
(after resolving redirects)		
Unidentified due to mislabelling the Country property as County	5	
Unidentified due to '.svg' file of the flag as value for the country	35	
Partially found with remaining as outliers	3	11
Partially found without outliers		7
Misaligned with a parent concept	8	6
Other problems	7	2

Out of the 63 country alignments detected, 26 were correct. 27 other alignments had a '.svg' file appearing as value of the country property in *DBpedia*. We would have detected such *concept coverings*, had such assertions for the country property been correct. Since this is a problem with the data and not our algorithm, we consider these 27 as correct for this particular evaluation. We thus get a precision of 84.13% ((26+27) out of 63). The two sources contained around 169 possible country alignments between them, including countries with a '.svg' value for the country property. There were many alignments in the ground truth that were not found because the system did not have enough support ($R < 0.9$) to pass our threshold. Accordingly, the recall was 31.36% and the F-measure was 45.69%.

Linking *LinkedGeoData* with places in *DBpedia*: We evaluated 200 alignments found between *LinkedGeoData* and *DBpedia*, out of which 157 were found to be correct, giving a precision of 78.2%. Common patterns of alignments include alignments of an area identified by its *OpenGeoDb* location id with its name or license plate codes from *DBpedia*. We were not able to detect 14 alignments correctly, where there were multiple spellings for the same entity (e.g. *LinkedGeoData* uses both "Hof Oberfranken" and "Landkreis Hof Oberfranken" in its values for its *linkedgeodata:is_in* property). Another 20 alignments evaluated were partial (e.g. out of the 88 counties in Ohio, the algorithm produced a covering including only 54). There were some other errors as well (e.g. Places with license plate code GR in *DBpedia* were aligned with instances having license code GR, NOL & ZI in *LinkedGeoData*).

Linking Species from *Geospecies* with *DBpedia*: In aligning *Geospecies* with *DBpedia*, out of the 178 alignments that we evaluated, we found 109 correct alignments for a precision of 61.24%. For 25 of the results, due to the presence of multiple names/lexical values for the same item (e.g. both"Decapoda"@en and dbpedia:Decapoda values exist for *dbpedia:ordo* property). In 28 of the evaluated alignments, we were only able to find partial *concept coverings*, mostly because of insufficient instances and property assertions. For 16 other alignments, however,

Table 3. Linking *LinkedGeoData* with *DBpedia*

Description of Pattern Observed	Alignments w/ larger class from *LinkedGeoData*	Alignments w/ larger class from *DBpedia*
Total # Alignments	2746	3097
# Alignments Evaluated	**100**	**100**
Correct	**78**	**79**
Unidentified due to multiple spellings	5	9
Partially found	13	7
Other	4	5

Table 4. Linking *Geospecies* with *DBpedia*

Description of Pattern Observed	Alignments w/ larger class from *Geospecies*	Alignments w/ larger class from *DBpedia*
Total # Alignments	191	255
# Alignments Evaluated	**93**	**85**
Correct	**49**	**60**
Unidentified due to multiple spellings	25	0
Partially found	4	24
Other	15	1

there were some interesting reasons. In some cases, the biological classes were no longer in use (Urticales, Homoptera, etc.). There were some alignments that we were not able to guess correctly because the species were marked as belonging to different classification systems. There were also a few mismatches to a class at a different level in the hierarchy.

Linking Genes from *GeneID* with *MGI*: In the 28 alignments found between *GeneID* and *MGI*, 24 were found to be correct for a precision of 85.71%. Most (20) of these were alignments linking a gene start position from *MGI* with possible locations from *GeneID*. In theory, these are numeric distances in centimorgans and can actually be an infinite set. In the data however we find all possible distances occurring as text. The other 4 alignments were partial because of insufficient data.

Table 5. Linking *GeneID* with *MGI*

Description of Pattern Observed	Alignments w/ larger class from *GeneID*	Alignments w/ larger class from *MGI*
Total # Alignments	6	22
# Alignments Evaluated	**6**	**22**
Correct	**4**	**20**
Partially found	2	2

5 Related Work

Ontology alignment and schema matching have been a well explored area of research since the early days of ontologies [5,1] and received renewed interest in recent years with the rise of the Semantic Web and Linked Data. In the Web of Linked Data, even though most work done 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. [8] describe the BLOOMS approach, which uses a central forest of concepts derived from topics in Wikipedia. An update to this is the BLOOMS+ approach [9] that aligns Linked Open Data ontologies with an upper-level ontology called Proton. BLOOMS is unable to find alignments because of the single *Feature* class in *GeoNames*. BLOOMS+, which uses contextual information, finds some alignments between *GeoNames* & Proton (precision of 0.5%) and *DBpedia* & Proton (90%). 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 achieves a precision of 26% and a recall of 68%. In comparison, for *GeoNames* and *DBpedia*, we achieve a precision of 64.4%. But this comparison does not reflect that we find concept coverings in addition to one-to-one alignments, while the other systems only find one-to-one alignments. 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* even in the presence of a rudimentary ontology. We believe that since other approaches do not consider concept descriptions beyond those in the original ontology (like *concept coverings*), they 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.

Extensional techniques and concept coverings have also been studied in the past [7]. Völker et al. [13] describe an approach, similar to our work, that uses statistical methods for finding alignments. This work induces schemas for RDF data sources by generating OWL-2 axioms using an intermediate associativity table of instances and concepts (called *transaction datasets*) and mining associativity rules from it. The GLUE [3] system is a instance-based matching algorithm, which first predicts the concept in the other source that instances belong to using machine learning. GLUE then hypothesizes alignments based on the probability distributions obtained from the classifications. Our approach, in contrast, depends on the existing links (in Linked Open Data Cloud), and hence reflects the nature of the source alignments in practice. *CS*R [12] is a similar work to ours that tries to align a concept from one ontology to a union of concepts from the other ontology using the similarity of properties 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.

Table 6. Example alignments from *GeoNames-DBpedia* *LinkedGeoData-DBpedia*

#	r_1	$p_2 = \{v_2\}$	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
***DBpedia* (larger) - *GeoNames* (smaller)**							
1	As described in Section 4, Schools, Colleges and Universities in *GeoNames* make Educational Institutions in *DBpedia*						
	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
2	The concepts for the country Spain are equal in both sources. The only outlier has its country as Italy, an erroneous link.						
	dbpedia:country = dbpedia:Spain	*geonames:countryCode = {ES}*	0.9997	3917	3918	IT (1/7635)	3918
3	We confirm the hierarchical nature of administrative divisions with alignments between administrative units at two different levels.						
	dbpedia:region = dbpedia:Basse-Normandie	*geonames:parentADM2 =* {geonames:2989247, geonames:2996268, geonames:3029094}	1.0	754	754		754
4	In aligning airports, an airfield should have been an an airport. However, there was not enough instance support.						
	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
***GeoNames* (larger) - *DBpedia* (smaller)**							
5	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*						
	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 (1/3)	1940
6	The error pattern in #5 seems to repeat systematically, as can be seen from this alignment for the country of Jordan.						
	geonames:countryCode = JO	*dbpedia:country =* {dbpedia:Jordan, dbpedia:Flag_of_Jordan.svg}	0.95	19	20		19
***DBpedia* (larger) - *LinkedGeoData* (smaller)**							
7	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. Instances supporting the remaining 2 license plates were missing						
	dbpedia:bundesland = Saarland	*lgd:OpenGeoDBLicensePlate-*Number = { HOM, IGB, MZG, NK, SB, SLS, VK, WND}	0.93	46	49		46

Table 7. Example alignments from *LinkedGeoData-DBpedia*, *Geospecies-DBpedia*

#	r_1	$p_2 = \{v_2\}$	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
8	Schools in *DBpedia* can be explained with types K2543 and Schools from *LinkedGeoData*						
	rdf:type= dbpedia:School	*rdf:type =* {lgd:School, lgd:K2543}	0.9907	2356	2378		2356
***LinkedGeoData* (larger) - *DBpedia* (smaller)**							
9	Due to missing instance links, this *concept covering* incorrectly claims that the state of New Jersey is composed of 9 counties while actually it has 21.						
	lgd:gnisST_alpha = NJ	*dbpedia:subdivisionName =* {Atlantic, Burlington, Cape May, Hudson, Hunterdon, Monmoth, New Jersey, Ocean, Passaic}	1.0	214	214		214
10	Waterways in *LinkedGeoData* is equal to the union of streams and rivers from *DBpedia*						
	rdf:type = lgd:Waterway	*rdf:type =* {dbpedia:River dbpedia:Stream}	0.97	33	34	dbpedia:Place(1/94989)	34
***DBpedia* (larger) - *Geospecies* (smaller)**							
11	Species from *Geospecies* with the order names Anura, Caudata & Gymnophionia are all Amphibians We also find inconsistencies due to misaligned instances, e.g. one amphibian was classified as a Turtle (Testudine) .						
	rdf:type = dbpedia:Amphibian	*geospecies:hasOrderName =* {Anura, Caudata, Gymnophionia}	0.99	90	91	Testudines (1/7)	91
12	Upon further inspection of #11, we find that the culprit is a Salamander						
	rdf:type = dbpedia:Salamander	*geospecies:hasOrderName =* {Caudata}	0.94	16	17	Testudines (1/7)	17
***Geospecies* (larger) - *DBpedia* (smaller)**							
13	The Kingdom Plantae, from both sources, almost matches perfectly. The only inconsistent instance happens to be a fungus.						
	rdf:type = dbpedia:Plant	*geospecies:inKingdom =* {geospecies:kingdoms/Ab}	0.99	1874	1876	geospecies:kingdoms/Ac (1/8)	1875

Table 8. Example alignments from *Geospecies-DBpedia* and *GeneID-MGI*

#	r_1	$p_2 = \{v_2\}$	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
14	Inconsistencies in the object values can also be seen - Carnivores from *Geospecies* are aligned with both Carnivora and Carnivore.						
	geospecies:inOrder = geospecies:orders/jtSaY	*dbpedia:ordo =* {dbpedia:Carnivora, dbpedia:Carnivore}	0.99	246	247		246
15	We can detect that species with order Chiroptera correctly belong to the order of Bats. Unfortunately, due to values of the property being the literal "Chiropta@en", the alignment is not clean.						
	geospecies:hasOrderName = Chiroptera	*dbpedia:ordo =* {Chiroptera@en, dbpedia:Bat}	1	111	111		111
***GeneID* (larger) - *MGI* (smaller)**							
16	The classes for Pseudogenes align.						
	bio2rdf:subType = pseudo	*bio2rdf:subType =* {Pseudogene}	0.93	5919	6317	Gene (318/24692)	6237
17	The Mus Musculus (house mouse) genome is composed of complex clusters, DNA segments, Genes and Pseudogenes.						
	bio2rdf:xTaxon = taxon:10090	*bio2rdf:subType =* {Complex Cluster/Region, DNA Segment, Gene, Pseudogene}	1	30993	30993		30993
***MGI* (larger) - *GeneID* (smaller)**							
18	Inconsistencies are also evident as the values pseudo and Pseudogene are used to denote the same thing.						
	bio2rdf:subType = Pseudogene	*bio2rdf:subType* = {*pseudo*}	0.94	5919	6297	other (4/230) protein-coding (351/39999) unknown(23/570)	6297
19	We find alignments like #19 & #20 , which align the gene start (with the chromosome) in *MGI* with the location in *GeneID* As can be seen, the values of the locations (distances in centimorgans) in *GeneID* contain the chromosome as a prefix. Inconsistencies 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.						
	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

6 Conclusions and Future Work

We described an approach to identifying *concept coverings* in Linked Data sources from the Geospatial, Biological Classification and Genetics domains. By introducing the 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 produces coverings where concepts at different levels in the ontologies of two sources can be mapped even when there is no direct equivalence or only rudimentary ontologies exist. Our algorithm is also able to find outliers that help identify erroneous links or inconsistencies in the linked instances. Our results provide a deeper insight into the nature of the alignments of Linked Data.

In future work we want to find 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. By mining rules from the generated alignments, we will be closer to the interoperability vision of the Semantic Web. A second direction of future work is to use the outliers to feed the corrections back to the sources, particularly *DBpedia*, and to the RDF data quality watchdog group pedanticweb.org. To achieve this satisfactorily, we not only need to point out the instances that have errors, but suggest why those errors occurred, that is, whether it was due to incorrect assertions or missing links.

Acknowledgements. This research is based upon work supported in part by the National Science Foundation under award number IIS-1117913. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or any person connected with them.

References

1. Bernstein, P., Madhavan, J., Rahm, E.: Generic schema matching, ten years later. Proceedings of the VLDB Endowment 4(11) (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. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Ontology matching: A machine learning approach. In: Handbook on Ontologies, pp. 385–404 (2004)
4. Duckham, M., Worboys, M.: An algebraic approach to automated geospatial information fusion. International Journal of Geographical Information Science 19(5), 537–558 (2005)
5. Euzenat, J., Shvaiko, P.: Ontology matching. Springer (2007)
6. Horrocks, I., Patel-Schneider, P., Van Harmelen, F.: From shiq and rdf to owl: The making of a web ontology language. Web Semantics: Science, Services and Agents on the World Wide Web 1(1), 7–26 (2003)

7. Isaac, A., van der Meij, L., Schlobach, S., Wang, S.: An Empirical Study of Instance-Based Ontology Matching. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ISWC/ASWC 2007. LNCS, vol. 4825, pp. 253–266. Springer, Heidelberg (2007)
8. 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)
9. 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)
10. 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)
11. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Finding concept coverings in aligning ontologies of linked data. In: Proceedings of the First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data in Conjunction with the 9th Extended Semantic Web Conference, Heraklion, Greece (2012)
12. 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)
13. 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)