Aligning Unions of Concepts in Ontologies of Geospatial Linked Data

Rahul Parundekar, José Luis Ambite, and Craig A. Knoblock

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

Abstract. It is evident from the recent growth in Geospatial Linked Data that even though the number of instances being generated and linked has increased drastically, the ontologies behind these sources remain disconnected. Though we can agree that the instances being linked are equivalent, the alignments that are extrapolated from these links between the concepts may or may not agree with our intuitions. It is important to investigate how the concepts in the sources are actually aligned. Our previous work was successful in finding alignments, such as equivalence and subset relations, between concepts of two sources, using the instances that are linked as equal. Such alignments need not be trivial, however, as a concept in the ontology might not have an exact equivalent class in the other source. In this paper we propose a method that uses the subset and equivalence relations between *restriction classes* found by our previous work to find new alignments, where one (larger) concept of a source is aligned to the union of multiple (smaller) concepts from another source. We also show that we can use these alignments to find inconsistencies and use them to identify the instances that may be erroneously aligned.

1 Introduction

The Web of Linked Data has seen huge growth in the past few years. As of September 2010, the size of the Linked Open Data Cloud was about 28.5 billion triples with around 20.6% of the triples belonging to the geospatial domain.[1] As of June 2009, the cloud had recorded an overall growth of about 300% with 91% growth in the geospatial domain.[2] Out of the 16 geospatial data sources covered in the September 2010 count, there are around 16.5 million outgoing links to other sources. The sources of Geospatial Linked Data are most popularly connected using the *owl:sameAs* property, linking instances that are the same. As more alignments are generated in the Web of Linked Data at the instance level, a pattern of inter-linked data arises where the ontologies behind the sources

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

[2] http://events.linkeddata.org/ldow2011/slides/ldow2011-slides-intro.pdf

remain un-linked. As described in our previous papers on Linking and Building Ontologies of Linked Data [7] and Aligning Ontologies of Geospatial Linked Data [6], an extensional technique can be used to generate alignments between the ontologies behind these sources. In these papers, we introduce a concept of *restriction classes*, which is similar to that of single value constraints on property restrictions of the *Web Ontology Language (OWL)* to increase the expressivity of sources with a rudimentary ontology. By looking at the set containment relationships of the instance sets of these *restriction classes*, we find equivalent and subset alignments between the two sources. Though the equivalent alignments found are precise in finding similar concepts between the two sources, the subset relations found, though informative, are too numerous to be effectively used.

Reviewing these subset relations we discovered that there are potential equivalent alignments not found by our previous work, linking a larger concept to a union or aggregation of one or more of its subsets. Using this as motivation, the work described in this paper builds on the ontology alignment method of [7]. Picking up where we left off, the approach described in this paper uses the subset relations as hints to create a union of smaller *restriction classes*, by virtue of a common property and *restriction classes* with only a single *property-value pair*, which guides the aggregation and then performs set containment operations with the larger *restriction class* from the other source. Using this method, we explore three Geospatial Linked Data sources - *GeoNames*, *DBpedia*, & *LinkedGeoData* and try to find *new* alignments between *GeoNames* & *DBpedia* and *LinkedGeoData* & *DBpedia*, where a larger subsuming *restriction class* from one source can be explained by an aggregation of smaller *restriction classes* from the other source.

The scope of this paper is in the domain of Geospatial Linked Data, where we find alignments between three sources: *GeoNames*, *DBpedia* and *LinkedGeoData*. We first find equivalences and subset relations as described in our previous work, and then use these to find the new *union alignments*. The nature of each of the three sources investigated is briefly mentioned here and they are described in more detail in [7]. *GeoNames* is a geographic source with a flat-file like ontology where all instances belong to a single concept of *Feature* and have associated *Feature Class & Feature Code* property to identify the instances as mountains, lakes, etc. Although *DBpedia* is a Linked Data source that covers domains other than the geospatial domain, there are a large number of instances from *GeoNames* linked to those in *DBpedia* using the *owl:sameAs* property. We also try to find alignments between the ontologies behind *LinkedGeoData* and *DBpedia*. RDF data in *LinkedGeoData* is derived from the *Open Street Map* initiative and has links to *DBpedia*.[3]

This paper is organized as follows. We first describe briefly our alignment algorithm from [7] along with the limitations of the results that were generated. We then explain our approach to finding alignments between a larger concept from one source and the union set of multiple smaller concepts from the other source. This is followed by identifying the outliers of these alignments that high-

[3] http://linkedgeodata.org/Datasets

light the inconsistencies and the instances that are erroneously linked. We then describe the experimental results that contain the new alignments discovered in these data sources, along with their outliers. Finally, we describe other related work and conclude with our observations and future work.

2 Aligning geospatial ontologies on the Web of Linked Data

The work described in this paper follows our previous work on aligning ontologies of Linked Open Data, which uses an extensional approach to find alignments between *restriction classes* in two different sources. Though the results generated by our previous algorithm found equivalent alignments between the two sources, a large number of subset alignments were also found. A pattern was observed in these results, where a group of concepts from one source were subsets of the same larger concept from the other source. In many cases these smaller concepts taken together were able to completely explain the larger source. We used this insight as motivation for consuming the subset relations, which were too numerous to be useful by themselves, to find alignments between the larger concept and the union of the group of concepts. Our approach uses this group of smaller concepts and introduces a disjunction operator on these subsets to try to define the common subsuming concept.

2.1 Our previous work on linking and building ontologies of Linked Data

Ontologies of Linked Data sources can be quite rudimentary. For example, *GeoNames* only has a single concept (*Feature*) to which all of its instances belong. On the other hand, in *DBpedia*, we find a rich ontology with a hierarchy of concepts and well-defined properties. In the traditional sense of ontology alignment, we would have found at most a single alignment between *Feature* on the *GeoNames* side and a similar broad concept from *DBpedia.* In order to get a richer set of alignments, we introduced the concept of a *restriction class.* A *restriction class* is a concept that is derived extensionally and defined by the set of instances obtained by restricting a single property to a single value (called a *property-value pair* and represented by $(p_i = v_i)$) in a source. For example, a *restriction class* for schools can be constructed in *GeoNames* by forming a set of instances that have their *geonames:featureCode* restricted to '*S.SCH*'. This *restriction class* is represented as *geonames:featureCode=S.SCH.* The scope of the definition of a *restriction class* includes the conjunction operator, which produces a more specialized set of instances, constructed using two or more *restriction classes.* Thus, a *restriction class* {*geonames:featureCode=S.SCH & geonames:countryCode=US*}, built from the *restriction classes geonames:featureCode=S.SCH* and *geonames:countryCode-=US*, can be defined by the intersection of the two sets and forms a concept extensionally described by the set of schools in the US in *GeoNames.*

Our algorithm aligns *restriction classes* from two sources, using an extensional technique, as follows. A pre-processing step first performs an inner-join on the two sources to be aligned based on an instance equivalence property like *owl:sameAs*. As inverse functional properties can only result in *restriction classes* with a single instance belonging to it, the pre-processing step eliminates them. The crux of the algorithm uses a top-down tree exploration of the space of alignment hypotheses. At the topmost level, a seed hypothesis is generated by aligning a *restriction class* with one *property-value pair* from the first source with another *restriction class* with one *property-value pair* from the second source. At each level in the search space, a new *restriction class* is formed from one *restriction class* of one of the sources by adding another *property-value pair* constraint on that *restriction class*. A new alignment hypothesis is thus constructed from the new *restriction class* and the *restriction class* from the other source. Each alignment hypothesis is tested for set containment relations between the intersection set of the *restriction classes* from both sources. This is done with the help of two scoring functions - P & R. If r_1 and r_2 are the two *restriction classes* in the alignment hypothesis, we first define Img(r_1) as the set of instances in the second source that instances of r_1 are linked to. We then define P as $\frac{|Img(r_1) \cap r_2|}{|r_2|}$, and R as $\frac{|Img(r_1) \cap r_2|}{|Img(r_1)|}$. We mark the relation of the alignment hypothesis as either i) equivalent ($P = 1, R = 1$), ii) subset, with the *restriction class* from the first source as extensionally subsuming the *restriction class* from second source ($R = 1$), iii) subset, with *restriction class* from second source extensionally subsuming the *restriction class* from first source ($P = 1$) or iv) no relation between the two *restriction classes*. To compensate for missing and misaligned instances, we relax our subset scores by defining P' and R' that reduce the required fraction of support to be greater than 0.9 instead of equal to 1. For an optimal exploration of the search tree, we employ certain pruning mechanisms that include i) using ordered exploration to avoid exploring a node twice, ii) pruning a node if the intersection set of the *restriction classes* of the hypothesis has size less than a minimum support size (we used 10 in our experiements), iii) pruning a node if the added *restriction class* does not change the set of instances, etc. After the brute-force exploration of the search space of alignment hypotheses, we use a post-processing step on the results generated, which removes redundant assertions by virtue of set containment of instances of two hypotheses where one is the immediate parent of the other in the search tree.

At the end of the above three steps of processing, the algorithm was able to find equivalent relations between *restriction classes* from two sources as well as subset relations in either direction. As this algorithm was not specific to any particular domain, we explored candidate sources for alignments in three domains: Geospatial, Genetics and Zoology. In these three domains, our algorithm found alignments of 5 pairs of sources. For example, we were able to find alignments between *GeoNames* and *DBpedia* in the Geospatial domain. One such alignment was the equivalent relation between {*geonames:countryCode=ES*} and {*dbpedia:country=Spain*} (i.e. correctly aligning the concepts for the country Spain). We also found subset relations like {*geonames:featureCode=S.SCH*} subset of *rdf:type=dbpedia:EducationalInstitution*. More such results are described in [7].

Limitations The approach above produced a large number of equivalent alignments that gave an exact mapping between the two *restriction classes* from the two sources. It also, however, produced a large number of subset relations that were not as useful. This was mainly because the subset relations, by themselves, did not contribute to a useful equivalence alignment between two classes. In all, in the *GeoNames* and *DBpedia* alignment, there were 1647 subset relations found. Though it is understandable that in many cases there might never exist an exact equivalence between two *restriction classes*, because they were auto-generated using *property-value pairs*, we decided to look for additional useful alignments, if any, that these subset relations might be able to provide us. For example, in the *GeoNames* and *DBpedia* alignment, we found that {*geonames:featureCode=S.SCH*}, {*geonames:featureCode=S.SCHC*} and {*geonames:featureCode=S.UNIV*} (i.e. Schools, Colleges and Universities from *GeoNames*) are all subsets of {*rdf:type=dbpedia:EducationalInstitution*}. Taken individually, though each of these alignments are correct and insightful, they are not particularly useful in understanding the relationships between *GeoNames* and *DBpedia*. Taken together, however, we found that the union of these three *restriction classes* completely define *rdf:type=dbpedia:EducationalInstitution*. The limitation of our approach was in the expressivity of our *restriction classes*. Though it included *restriction classes* containing single *property-value pairs* and the conjunction operator on those *restriction classes*, it did not include a disjunction operator and hence was unable to make use of the subset relations.

2.2 Identifying spatial concept coverings

As explained above, we were able to identify a pattern where a group of *restriction classes* from one source were aligned as subsets of a common concept from the other source. By using these alignments as hints, we were able to construct the union of the smaller *restriction classes* and detect if the union was able to define the larger class entirely. The following section describes this method in detail. In those cases where we are not able to define the larger class entirely, our approach is also able to find and explain the missing instances (*outliers*).

Mapping a *restriction class* from one source with a union of smaller *restriction classes* from the other source Since the problem of finding alignments with conjunctions and disjunctions of *property-value pairs* of *restriction classes* is combinatorial in nature, we focus only on subset relations where both *restriction classes* have a single *property-value pair* and where one is a subset of the other. 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*. We then identify a strategy for selecting the smaller *restriction classes* from within such a group to form the union that best describes the larger *restriction class*. 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 *restriction classes* represents the union of their set of instances, *ii)* each of the *restriction classes* that are aggregated contain only a single *property-value pair* and *iii)* the property is the same for all those *property-value pairs*. We then try to find the alignment between the larger common *restriction class* and a set of *restriction classes* from the other source that are aggregated by the disjunction operator by using an extensional approach similar to our previous paper. We call such an alignment as *union alignment*.

We first build candidates for aggregation using the results from our previous algorithm as hints. We group alignments by the larger common *restriction class*. Grouping the subset relations is trivial. Equivalence relationships are subsets in both directions and thus are easily integrated into the groups. For each alignment, $\{p_1{=}v_1\}$ is the r_1 part and $\{p_2{=}v_2\}$ forms the r_2 part (each with a single *property-value pair*) as explained in the previous section. Sub-groups are formed by aggregating according to the property of the *property-value pairs* of the smaller *restriction classes*. Such a sub-group is identified by {*Property of the larger restriction class*(p_1)*, Value of the larger restriction class*(v_1)*, property of the smaller restriction classes*(p_2)}. Values of the different smaller *restriction classes* can be denoted by a list *List*(v_2s)*.* The disjunction of the smaller *restriction classes*creates a set of instances that extensionally identifies the union concept. We can now either confirm or refute the hypothesis that the larger *restriction class* is equivalent to the union concept. We can do this by using a scoring mechanism similar to the use of P & R in our previous paper. Using the same terminology, U_A is defined as the set of disjunctive instances (i.e. Union($Img(r_1) \cap r_2$))), U_L is defined as the set of instances of the larger class taken by itself (i.e. Img(r_1)) and U_S is defined as the set of instances that is the union of individual smaller *restriction classes*(i.e. Union(r_2)). The scoring mechanism defines P_U as $\frac{U_A}{U_S}$ and R_U as $\frac{U_A}{U_L}$. P'_U & R'_U are defined as fractions with relaxed scoring assumptions similar to P' & R' from our previous paper.

For example, our previous algorithm finds that {*geonames:featureCode = S.SCH*}, {*geonames:featureCode = S.SCHC*}, {*geonames:featureCode = S.UNIV*} are subsets of {*rdf:type=dbpedia:EducationalInstitution*}. In this case, the subgroup can be identified as {*rdf:type, dbpedia:EducationalInstitution, geonames:featureCode*} and list as *(S.SCH, S.SCHC, S.UNIV).* As can be seen in the Venn diagram of Figure 1, U_L is the *restriction classImg(*{*rdf:type = dbpedia:EducationalInstitution*}*),* U_S is {*geonames:featureCode = S.SCH*} ∪ {*geonames:featureCode = S.SCHC*} ∪{*geonames:featureCode = S.UNIV*} and U_A is:

{Img({rdf:type = dbpedia:EducationalInstitution}) ∩ {geonames:featureCode = S.SCH}} ∪ {Img({rdf:type = dbpedia:EducationalInstitution}) ∩ {geonames:featureCode = S.SCHC}} ∪ {Img({rdf:type = dbpedia:EducationalInstitution}) ∩ {geonames:featureCode = S.UNIV}}

Ideally, for an exact equivalence alignment, P'_U & R'_U should both be 1.0, if the larger *restriction class* covers the union of the smaller *restriction classes* completely and vice-versa. However, similar to the relaxed score assumption from our previous paper to accommodate errors in the dataset, we consider it a complete coverage when the score is greater than a relaxed score of 0.9. (i.e. the *union alignment* is considered to be equivalent if $P'_U > 0.9$ & $R'_U > 0.9$). Due to the minimum support score constraint for subsets from our previous paper, we are assured that $\frac{U_A}{U_S}$ i.e. P'_U is always going to be greater than 0.9.[4] Thus, we can say that a *union alignment* is equivalent if $R'_U > 0.9$. 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. The scores for other *union alignments* found are described in the results section.

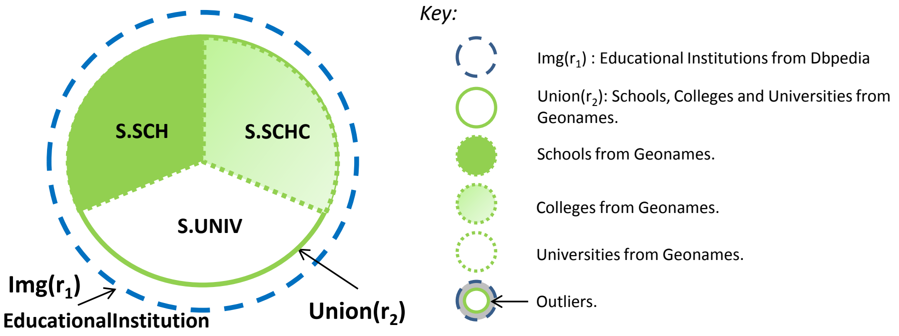
S.SCH
S.SCHC
S.UNIV
Img(r₁)
EducationalInstitution
Union(r₂)
Key:
Img(r₁) : Educational Institutions from Dbpedia
Union(r₂): Schools, Colleges and Universities from Geonames.
Schools from Geonames.
Colleges from Geonames.
Universities from Geonames.
Outliers.

Fig. 1. Spatial covering of Educational Institutions from *DBpedia*

Using mappings to identify outliers As mentioned above, the score for the alignment of $\{$*rdf:type = dbpedia:EducationalInstitution*$\}$ to the union of $\{$*S.SCH, S.SCHC & S.UNIV*$\}$ is approximately 0.98. For $\{$*rdf:type = dbpedia:EducationalInstitution*$\}$, 396 instances out of the 403 Educational Institutions were accounted for as having their *geonames:featureCode* as one of *S.SCH, S.SCHC or S.UNIV* to give this score. An interesting question to pose then is, how are the remaining 2% of the *dbpedia:EducationalInstitutions* (i.e. 7 instances) classified in *GeoNames*?

While calculating the disjuncted *restriction classes*, we also keep track of other instances with the same $\{p_1, v_1, p_2\}$ but not previously considered as subsets. These had been pruned in the exploration stage as they either had a size of less than the minimum support size constraint of ten instances or had P' less than 0.9. For the first type of *restriction classes*, those with low support size but yet having P' greater than 0.9 are now re-classified as subsets. The

[4] It should also be noted that each of the smaller subsets also satisfy the minimum support size of 10 instances.

re-classification of the relation as a subset can now be justified due to increased evidence in suggesting subsumption as other values for the same property are also aligned as subsets of the larger *restriction class* from the first source.

The second type of *restriction classes* that had P' less than 0.9 along with the ones that were not re-classified above (i.e. with less than 10 instances and P' less than 0.9) form the *outliers*. For example, as mentioned before, schools, colleges and universities from *GeoNames* make up 396 out of 404 Educational Institutions from *DBpedia*. From the other eight instances, 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). The eighth instance does not have a *geonames:featureCode* property asserted. The P' score of these *restriction classes* is less than 0.9. One of the instances classified as *dbpedia:EducationalInstitution* in *DBpedia* is linked to an instance in *GeoNames* that has *geonames:featureCode* as 'S.HSP'. [5] There are 31 instances in {*geonames:featureCode=S.HSP*}, however, and because this *restriction class* does not meet the relaxed subset score threshold, it cannot be considered in the union of *restriction classes*. Another example of outliers was found in the {*dbpedia:country = Spain* ≡ *geonames:countryCode = ES*} alignment. This equality was found using the relaxed subset assumption, where 3917 of the 3918 instances of *dbpedia:country=Spain* were accounted for as having *geonames:countryCode=ES*, resulting in a subset score of 0.9997. The one instance not having country code ES was actually classified as having country code IT (Italy). This single instance needs to be inspected further and it needs to be determined if the *owl:sameAs* link is correct. It is evident from the above examples that the outliers help in understanding the nature of the sources more explicitly, showing why the alignments failed to completely describe the larger *restriction class*. These, along with a few other examples, are described in detail in the next section.

3 Experimental Results

From the approach described in Section 2.2, we were able to get a total of 752 union alignments for the *GeoNames-DBpedia* alignment and 5843 for the *LinkedGeoData-DBpedia* alignment. From the 752 in *GeoNames-DBpedia*, 318 are such that the larger *restriction class* is from *DBpedia*, while the other 434 have the larger *restriction class* from *GeoNames*. Similarly, 3097 from the 5843 *union alignments* in *LinkedGeoData-DBpedia* have the larger *restriction class* from *DBpedia*, while the other 2746 have the larger *restriction class* from *GeoNames*. Tables 1, 2, 3, & 4 list a few interesting examples of these *union alignments* between *GeoNames-DBpedia* and *LinkedGeoData-DBpedia* (in either direction), which we describe here. The tables are organized as follows. Column 2 describes the sub-group, i.e. (p_1,v_1,p_2). Column 3 contains the list of the value part of the *property-value pairs* in the *restriction classes* of the smaller sets (i.e. List(v_2)). The score of the union is noted in column 4 ($R'_U = \frac{|U_A|}{|U_L|}$) followed by $|U_A|$ and

[5] Intuitively, it would make sense to the reader that this instance might perhaps be a hospital of a medical school.

$|U_L|$ in columns 5 and 6. Column 7 describes the outliers, i.e. values of v_2 that form *restriction classes* that aren't direct subsets of the larger *restriction class*. Each of these values also has a fraction with the number of instances that do belong to the larger *restriction class* of the total number of instances of the *restriction class* (or $\frac{|Img(r_1)|}{|r_2|}$). 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 we were able to explain using U_A and the outliers. For example, the *union alignment* of #1, 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). The seven of the eight outliers found (S.BLDG, S.EST, S.LIBR, S.MUS, S.HSP) are mentioned along with their P' fractions in column 7.

We also found some other interesting alignments. #2 shows the details of the Spain example mentioned briefly in Section 2.2. #3 shows a union alignment that aligns smaller sets or parts from *GeoNames* to a complete set. The region of Basse-Normandie in France is made up of three departments. The *restriction classes* of these three regions are constrained by the *geonames:parentADM2* property. #4 shows that Airports and Airbases make up 99% of the airports in *DBpedia*. From its outliers, one might argue that Airfields (S.AIRF) should also be included, but it was not as its P' score was lower than the threshold. Outliers also show that there is a Hill in geonames that has been classified as an airport. Even though this instance may be an airport in the hills, ontologically it doesn't make sense that a hill can be an airport. A similar case is observed in #8 where we find that there is at least one water tower in *LinkedGeoData* that is aligned with an Educational institution in *DBpedia*.

The *union alignment* #5 should have been as straightforward as alignment #2. Our approach was able to detect a pattern, however, that might have been overlooked after looking at individual instances. Netherlands from *GeoNames*, for example, should be aligned with the country Netherlands from *DBpedia*. 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* generated while importing Wikipedia data into *DBpedia* (the error seems systematic, see Jordan in #6).

Alignment #7 was able to explain 8 of the 10 license plate codes in the state (bundesland) of Saarland[6]. The ones that it missed were Ottweiler (OTW) and the police vehicle codes (SAL). Since the vehicle code SAL is not associated with any populated places in Saarland, it is quite possible that it does not get mentioned in *LinkedGeoData*. Our approach thus provides a deeper insight into the nature of the sources. #9 tries to find the composition of the state of New Jersey. 100% of the instances in New Jersey from *LinkedGeoData* can be accounted

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

for in the 9 counties. New Jersey actually has 21 counties[7]. This suggests that instances in New Jersey in *LinkedGeoData* that are linked to *DBpedia* are not a complete representation resulting in an equivalent alignment. The quality of the results generated by our extensional approach are tied to the quality of the instances in the dataset. We find, however, that such alignments, even though they might be partially incorrect, give an accurate representation of the actual instances in the dataset and highlight the practical quality of the links in the Web of Linked Data.[8] Finally, alignment #10 describes how the concept Waterways in *LinkedGeoData* can be defined as the union concept of Streams and Rivers in *DBpedia*. The complete set of alignments discovered by our algorithm are available on our group page.[9]

4 Related Work

Ontology alignment has been a well explored area of research since the early days of ontologies. It has received renewed interest in recent years with the rise of the Semantic Web. Euzenat & Shvaiko [3] provide a comprehensive dis-usssion on Ontology Matching approaches. A closely related area of study to ontology alignment is schema matching. Bernstein et al. [1] summarize the developments in this field in the past ten years. 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 sources 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. Though we employ a simple set subsumption technique to identifying alignments, our use of *restriction classes* is able to find a large set of alignments in cases like aligning *GeoNames* with *DBpedia* or Proton, while BLOOMS & BLOOMS+ are unable to find alignments because of the small number of classes in *GeoNames* that have vague declarations. 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. Building ontologies of Linked Data sources using a statistical method has also been described in Völker et al. [8]. This work induces schemas for RDF data sources by generating OWL 2 axioms using intermediate associativity table of instances and concepts (called *transaction datasets*) and mining associativity rules from it.

[7] http://en.wikipedia.org/wiki/List_of_counties_in_New_Jersey

[8] In [7] we compared the extensional versus intensional perspective on ontology alignment. In a nutshell, the extensional alignment gives a precise characterization of the current relationship between the data in the sources, regardless of the intended meaning of the concept definitions. For example, a source may define instances as universities, but linkage can show that it only contains American universities.

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

Table 1. Example alignments from the *GeoNames* and *DBpedia* datasets, with larger sets from *DBpedia* and smaller sets from *GeoNames*

#	***Sub-group*** $\{p_1, v_1, p_2\}$	***List***(v_2)	$R'_U = \frac{\|U_A\|}{\|U_L\|}$	$\|U_A\|$	$\|U_L\|$	***Outliers***	***# Explained Instances***
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
2	{*dbpedia:country,* *dbpedia:Spain,* *geonames:countryCode*}	ES	0.9997	3917	3918	IT (1/7635)	3918
3	{*dbpedia:region,* *dbpedia:Basse-Normandie,* *geonames:parentADM2*}	geonames:2989247, geonames:2996268, geonames:3029094	1.0	754	754		754
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

Table 2. Example alignments from the *DBpedia* and *GeoNames* datasets, with larger sets from *GeoNames* and smaller sets from *DBpedia*

#	***Sub-group*** $\{p_1, v_1, p_2\}$	***List***(v_2)	$R'_U = \frac{\|U_A\|}{\|U_L\|}$	$\|U_A\|$	$\|U_L\|$	***Outliers***	***# Explained Instances***
5	{*geonames:countryCode,* *NL,* *dbpedia:country*}	dbpedia:Netherlands, dbpedia:The_Netherlands, dbpedia:Flag_of_the _Netherlands.svg	0.9802	1939	1978	dbpedia:Kingdom_of _the_Netherlands	1940
6	{*geonames:countryCode,* *JO,* *dbpedia:country*}	dbpedia:Jordan dbpedia:Flag_of_Jordan.svg	0.95	19	20		20

Table 3. Example alignments from the *LinkedGeoData* and *DBpedia* datasets, with larger sets from *DBpedia* and smaller sets from *LinkedGeoData*

#	***Sub-group*** $\{p_1, v_1, p_2\}$	***List***(v_2)	$R'_U = \frac{\lvert U_A\rvert}{\lvert U_L\rvert}$	$\lvert U_A\rvert$	$\lvert U_L\rvert$	***Outliers***	***# Explained Instances***
7	{*dbpedia:bundesland,* *Saarland,* *lgd:OpenGeoDBLicensePlateNumber*}	HOM, IGB, MZG, NK, SB, SLS, VK, WND	0.93	46	49		46
8	{*rdf:type,* *dbpedia:EducationalInstitution,* *rdf:type*}	lgd:Amenity, lgd:K2543, lgd:School, lgd:University, lgd:WaterTower	0.9901	2609	2610		2609

Table 4. Example alignments from the *LinkedGeoData* and *DBpedia* datasets, with larger sets from *LinkedGeoData* and smaller sets from *DBpedia*

#	***Sub-group*** $\{p_1, v_1, p_2\}$	***List***(v_2)	$R'_U = \frac{\lvert U_A\rvert}{\lvert U_L\rvert}$	$\lvert U_A\rvert$	$\lvert U_L\rvert$	***Outliers***	***# Explained Instances***
9	{*lgd:gnisST_alpha,* *NJ,* *dbpedia:subdivisionName*}	Atlantic, Burlington, Cape May, Hudson, Hunterdon, Monmoth, New Jersey, Ocean, Passaic	1.0	214	214		214
10	{*rdf:type,* *lgd:Waterway,* *rdf:type*	dbpedia:Stream, dbpedia:River	0.97	33	34	dbpedia:Place(1/94989)	34

5 Conclusions and Future Work

We described an approach to identifying *union alignments* in geospatial data sources on the Web of Linked Data. By extending our definition of *restriction classes* with the disjunction operator, we were able to find alignments of union concepts from one source to larger concepts from the other source. Our approach produced *union alignments* as results that found that concepts at different levels in the ontologies of two sources can be mapped even when there was no direct equivalence. We were 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 Geospatial Linked Data.

Though the scope of this paper is the geospatial domain, our algorithm can be used in other domains as well. Our next step is to explore other domains like zoology and genetics for *union alignments*. Other possible future work is in the mapping and understanding of the properties in 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.

Acknowledgements

This research is based upon work supported in part by the National Science Foundation under award number IIS-1117913.

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. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag (2007)
4. Jain, P., Hitzler, P., Sheth, A., Verma, K., Yeh, P.: Ontology alignment for linked open data. The Semantic Web–ISWC 2010 pp. 402–417 (2010)
5. Jain, P., Yeh, P., Verma, K., Vasquez, R., Damova, M., Hitzler, P., Sheth, A.: Contextual ontology alignment of lod with an upper ontology: A case study with proton. The Semantic Web: Research and Applications pp. 80–92 (2011)
6. 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)
7. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Linking and building ontologies of linked data. In: Proceedings of the 9th International Semantic Web Conference (ISWC 2010). Shanghai, China (2010)
8. Völker, J., Niepert, M.: Statistical schema induction. The Semantic Web: Research and Applications pp. 124–138 (2011)