A Gazetteer for Biodiversity Data as a Linked Open Data Solution

Silvio D. Cardoso[1], Kleberson J. Serique[1], Flor K. Amanqui[1]
J. L. Campos dos Santos[2], Dilvan A. Moreira[1]
[1]University of São Paulo, ICMC, São Carlos, Brazil
[2]National Institute for Amazonian Research, Manaus, Brazil
Email: {flork, serique}@icmc.usp.br, {silvio.domingos.cardoso, dilvan}@usp.br
{lcampos}@inpa.gov.br

***Abstract*—Biodiversity studies all life forms that we find in nature. The maintenance of biological diversity is important because it is essential to life on Earth. The lack of accurate spatial geographic information in species occurrence data, especially from diversity rich regions (like the Amazon Forest), leads to problems in many conservation activities, such as systematic planning for the protection of endangered species. In this paper, we present a gazetteer (a geographical directory that associate name places to geographic coordinates) for biodiversity data that is available as an Linked Open Data resource (using a GeoSPARQL Endpoint) and show how it can be used to improve inaccurate geographic collection data. We compared the efficiency of our Gazetteer with three openly available resources, Geonames, WikiMapia and Wikipedia, and got a 10% better recall rate than these endpoints. We also used the Gazetteer to correct geographic data from a big record sample (327,000 occurrence records) from SpeciesLink and GBIF (two big open access repositories of biodiversity occurrence data). In this data set, we were able to add geographic coordinates to around 14% of records that did not have them before.**

***Index Terms*—Biological Gazetteer; Geographic information retrieval; semantic web; SPARQL and GeoSPARQL.**

I. Introduction

Nowadays, there is great interest in biodiversity because it affects important economic activities [1], such as fishing, agriculture, and forestry and it is also an important factor for ecosystems' health. Therefore, several studies about biodiversity data access and recovery have been discussed within the academic community [2].

Biodiversity Data recovery can be performed in many repositories using Web applications. Leading biodiversity institutions, such as the New York Botanic Gardens, Smithsonian Institute and the National Institute of Amazonian Research, provide the data available in these repositories. But this data is difficult to analyze because it lacks a well-defined structure, uses specific biology vocabulary, and has spatiotemporal parameters, among other problems [2].

Using information about Biodiversity Informatics from the literature [3], it is possible to highlight the main challenges faced when analyzing this kind of data: (i) deal with large volumes of information, (ii) achieve interoperability of information from different sources and formats, (iii) manipulate data and images, (iv) handle geographic information.

Data from biodiversity repositories usually have a large number of records with inaccurate geographical information [4]. The lack of geographic spatial information accuracy in biodiversity data entails problems such as, the impossibility to propose sets of accurate locations for the protection of endangered species [5].

In this paper, we address the Biodiversity Informatics challenge of (i) and (iv). Our goal is improve the accuracy from geographic information using some techniques from the Geographic Information Retrieval (GIR) to improve the accuracy from geographical information contained in biological data and store them in a Gazetteer.

A Gazetteer is a geographical directory that associate name places to geographic coordinates. They are commonly implemented as directories that contain triples of place names (N), feature types (T) for named geographic places, and geographic footprints (F) with geographic coordinates [6]. They offer functions to map place names to footprints (N - F) and place names to feature types (N - T). Gazetteers are important for allowing geospatial queries, such as rivers in Washington State, to be performed by GIR systems [6].

In this paper, we show that data from two large biodiversity repositories, the Global Biodiversity Information Facility (GBIF)[1] and the SpeciesLink[2], have a large number of records with inaccurate geographical information. We have chosen these repositories because GBIF is a globally recognized resource and SpeciesLink have the support of several Brazilian institutions.

We also describe the development of a gazetteer, based on Linked Open Data technologies, for biodiversity data. Using this gazetteer we mapped geographical data from GBIF and Species to RDF triples and insert it in a GeoSPARQL endpoint. This kind of endpoint is capable of answering semantic queries. The advantage of providing Gazetteer data in a semantic format is that it makes it possible to perform complex queries, not possible with the structured data common in Gazetteers [7], for instance, retrieve all farms that border a forest reserve or return all specimens found to the north of a

[1]http://www.gbif.org/
[2]http://splink.cria.org.br/

978-1-4799-4249-7/14 $31.00 © 2014 IEEE
DOI 10.1109/WETICE.2014.19

city.

To evaluate our work, we selected a sample of queries to be used against three popular geography repositories, Geonames, Wikipedia and WikiMapia, and in our Gazetteer to compare their precision and recall. Our gazetteer got around 10% better recall and 8% better precision, when compared with others geography repositories.

We also used the gazetteer to recover geographical coordinates of 14% and 10% from GBIF and SpeciesLink records that did not have them before. This shows that we can effectively improve accuracy of biodiversity data.

II. Related Work

After a literature search about GIR systems that could be used for the construction of Gazetteers for biodiversity data, we found the solutions proposed by [3] [8] and [9].

The work proposed in [3] is a geographic annotation service for biodiversity systems. It uses data about the location of butterfly traps to make a connection between the collection data and the places where the specimens were found. This system implements a controlled environment where geographical information is accurate and available on the web[3] .

The gazetteer proposed in [8] has an approach based in a content expansion of the hydrographic data names of the Brazilian National Water Agency, which contains 5384 river names and 670 dikes. This gazetteer improves geographic data quality for research in biodiversity, however its data is not available on the web.

The system proposed in [9] is a gazetteer with urban places. It uses news texts to retrieve name places to improve the gazetteer data. However its data is stored in a relational database, is not specific to biodiversity data and is not available on the web.

As our literature search shows, there are not many GIR systems being used for the construction of gazetteers. Our gazetteer differs from these three systems because it does not restrict to one kind of specimen. (like [3]), it is specific to biodiversity data and it is openly available in the Web as a LOD (Linked Open Data) resource through a GeoSPARQL endpoint. Basically, LOD is a term that defines the best techniques for exposing, sharing, and connecting data, information and knowledge in the semantic web. Finally, because it uses GeoSPARQL, our gazetteer can perform complex geographical queries, not possible with the structured data in other Gazetteers.

III. Data Used

To evaluate our gazetteer, we used a sample of biodiversity collection data from the GBIF and speciesLink repositories for the Amazonas State in Brazil (downloaded in February 2014). We verified that around 42% of all data about Brazil in the GBIF repository had georeferenced data, for the speciesLink site this number was around 45.40%. These numbers show the lack of geographical information in Brazilian biodiversity data.

[3]http://www.lis.ic.unicamp.br/projects/biocore/

TABLE I: Quality metrics of GBIF site.

Geographical Information	Data quality	%	Information Quality
Place, Latitude, Longitude and county	25687	16.60%	4
Only place and county	11287	7.3%	3
Only place	98954	64.0%	2
Only county	5915	3.8%	1
Have no information	12886	8.3%	0

To evaluate the accuracy of spatial information in the data, we created a quality metric, shown in Table 1 and Table 2. Depending on how much geographic information a record has, it is classified in a level from 0 to 4. Only 16.60% of GBIF records and 24.85% of SpeciesLink records for the Amazonas State have accurate geographical information, i.e., contains the place name, latitude, longitude and municipality where a specimen was collected (level 4).

In GBIF and SpeciesLink data, around 31.56% of records ranking in quality levels 2 and 3 are from very old collections, between the years 1850-1979. At that time, GPS devices were not available and collections were not georeferenced (given the great efforts needed to find the coordinates of collection sites). Thus, they do not have accurate geographical information. This fact is discussed in [10], which shows the template used to collect biological data.

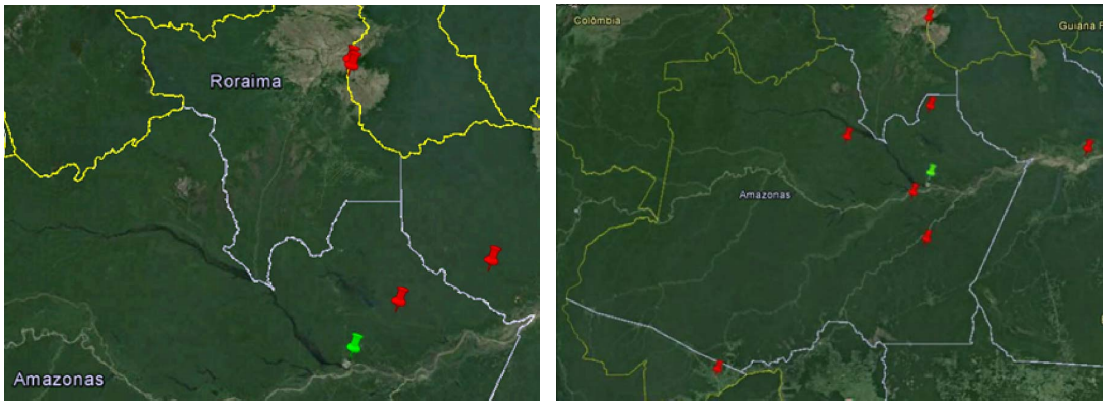
TABLE II: Quality metrics of SpeciesLink site.

Geographical Information	Data quality	%	Information Quality
Place, Latitude, Longitude and county	60786	24,85%	4
Only place and county	91419	37,4%	3
Only place	43961	18.0%	2
Only county	41071	16,8%	1
Have no information	7310	3%	0

Using our Gazetteer, we improved the quality of geographical information from records in levels 2 and 3 by referencing place names of records in level 4. Records in levels 2, 3 and 4 correspond to 87.9% of total records in GBIF and 80.2% in SpeciesLink. Records from levels 0 and 1 were discarded because they were too inaccurate, 12.1% of the records in GBIF and 20% in SpeciesLink were discarded.

We also observe in this sample that various localities have inaccurate latitude and longitude information, as shown in Figures 1 to 3. Figures 1a and 1b show locations for the Adolpho Ducke Reserve, the correct location of the reserve is shown as a green pin and the inaccurate locations as red pins.

Figures 2 and 3 show the specimen's location data for the whole sample. Given that the sample is only for the Amazonas State, there should not be collections outside its borders. As the figures show, that is not the case. The sample has location data showing specimens collected at sea, in Argentina, in Brazil neighboring countries (e.g., Venezuela, Colombia, Peru) and



(a) Coordinates for the Adolpho Ducke Reserve contained in GBIF data.

(b) Coordinates for the Adolpho Ducke Reserve contained in SpeciesLink data

Fig. 1: Coordinates for Adolpho Ducke Reserve contained in GBIF (a) and SpeciesLink (b) data. The red pins represent wrong geographic coordinates for the reserve. The green pin represents the correct coordinate for the reserve.

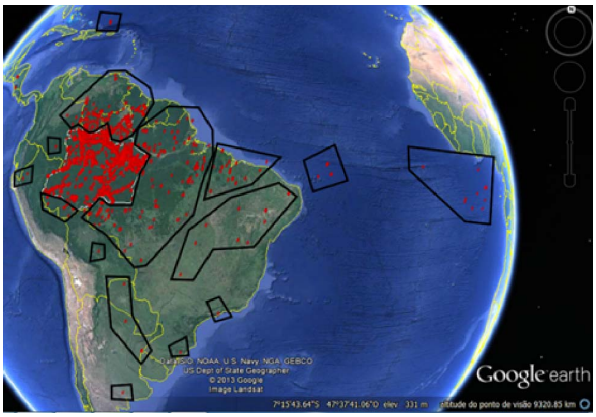


Fig. 2: Distribution of latitude and longitude coordinates for GBIF data.

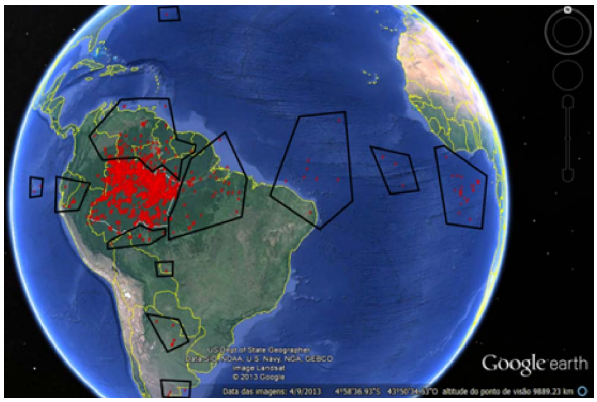


Fig. 3: Distribution of latitude and longitude coordinates for SpeciesLink data.

other states.

As can be seen, the sample data has two main problems with location data associated with specimen collection, some records have no latitude and longitude data and others have wrong (or inaccurate) coordinates. Our gazetteer can be used to correct some of this data.

A part from the sample from GBIF and speciesLink, we also used external source data from repositories, like Geonames, WikiMapia and Wikipedia to improve the accuracy of the data from INPA and GBIF. We chose these particular repositories because they are open and have a large number of popular places.

IV. Gazetteer Implementation

We followed the following steps in the Gazetteer development:

(i) Creation of a script for analyzing and processing GBIF and speciesLink data to remove invalid and low quality data, using the criteria shown in Table 1 and Table 2.

(ii) Clustering program that uses GIR techniques, such as toponym resolution (expressions used in place names) and place name disambiguation [11], [9].
For toponym resolution, we used a list composed by 45 place names in a XML file. Each place name contains a regular expression to retrieve information for a given a place, for instance the expression `(?i)reserva\b(.+?)[,|.|;|:]` retrieves all triples containing the word "reserva" in the beginning. We used the Star algorithm [12] together with a stop word list for clustering the places retrieved.
We also used the Geonames database as an external data source for place names disambiguation (whenever possible).

(iii) Analysis to improve geographical coordinates in the Gazetteer.
In this step, we verify the similarity between place

TABLE III: Clustering results to GBIF and SpeciesLink data.

Threshold	Number of distinct places	Accuracy of the groups created
0.4	4602	87.77%
0.5	4782	87.97%
0.6	4921	87.69%

TABLE IV: Sample values from selected centroids that increased the number of places.

- Reserva Florestal Adolpho Ducke.
- Reserva Florestal Ducke (Associação) Ha A3 próximo a Estrada.
- Reserva Florestal Duckec estr do Acará.
- Reserva Florestal Ducxke Manaus-Itacoatiara km 26 Área do Acará Floresta de Campinarana.

names and clustered groups. For that, we use the Jaccard similarity coefficient [13] with thresholds between 0.4 – 0.6, to evaluate the best range to be used in our Gazetteer. We selected a sample of 100 random groups to evaluate our technique. Table 3 shows that the accuracy of these groups had a low variation.

However, we observed that several groups were created for the same place, increasing the number of distinct places. The reason is that place names can have different spellings, ranging from misspellings to describe localities to different descriptions for the same places, as shown in Table 4.

This problem was mitigated with the use 0.4 as threshold. Even though a few place names remained as separate places, that does not affect the Gazetteer a lot as those place names still have coordinates located near each other.

For the problem of geographic coordinate inaccuracy, we created a method to improve the accuracy of such coordinates: It selects records and summarizes the geographical coordinates in a centroid coordinate. Values that appear more frequently in the group are replaced for all data, as shown in Figure 4.

This approach was chosen because the value, which occurs most frequently in a group, tends to be the correct value for a place. It guarantees around 87% accuracy for geographical coordinates, as show in Table 3.

(iv) Deployment of the Gazetteer as a GeoSPARQL endpoint.

In this step, we link the ontologies DBpedia, Linked GeoData and GeoSPARQL. The link between the GeoSPARQL ontology and the DBpedia is show in the Figure 5. For correct geographical inference, the GeoSPARQL ontology should be connected with others ontologies using its Feature class.

The GeoSPARQL ontology is important because it follows the patterns proposed by the Open Geospatial Consortium (OGC). OGC is an international industry consortium of 472 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards. Linking these ontologies with GeoSPARQL promotes an easier interchange of data using LOD technologies.

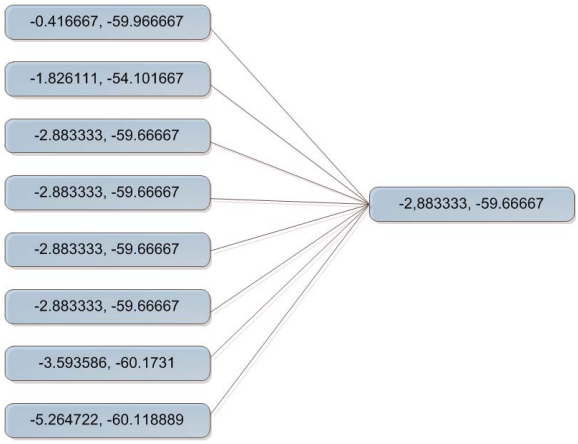


Fig. 4: Summarization of geographical coordinates. The numbers show geographic coordinates in GBIF and SpeciesLink data

After linking these ontologies, we map the Gazetteer data to terms on them, using a script, and store it as triples in a triple store. Figure 6 shows an example connecting terms from the GeoSPARQL ontology to Gazetteer data. The information in red represents the records present in the Gazetteer. The information in green shows the connection to a set of coordinates, using the `hasGeometry` property (from the GeoSPARQL ontology).

This connection is what makes it possible, for a GeoSPARQL enabled triple store, to perform geographical/topological inferences (such as be inside or border something). The geographical coordinate is represented by the wktLiteral property, as shown in the line 10 (Figure 6), "*geow:wktLiteral*".

The triple store used in our Gazetteer is the Parliament software [14]. We use it because it is a high-performance triple store, SPARQL/GeoSPARQL endpoint and reasoner.

V. Results and Discussion

To evaluate our Gazetteer, we used the sample data from GBIF and speciesLink. In that sample, we checked the amount of records that had latitude and longitude data, recovered after using the data summarization technique (discussed in section 4). GBIF had 16.60% of records with geographic coordinates (longitude and latitude) and SpeciesLink had 24.85%. It is important to highlight that a lot of these records are older occurrences, from the time GPS equipment was not available, being, for their age, invaluable.

After using our Gazetteer to add geographic coordinates to records without them, we obtained the results shown in Figure 7. The number of records with geographic coordinates in GBIF

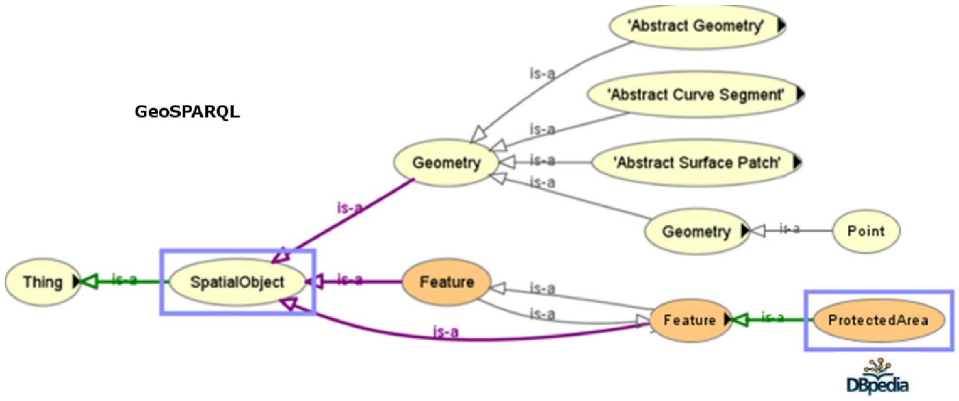


Fig. 5: GeoSPARQL ontology connection with other ontologies

```
1. <!– Gazetteer:Lake713 -->
2. <owl:NamedIndividual rdf:about="&Gazetteer;Lake713">
3. <rdf:type rdf:resource="dbp:Lake"/>
4. <Gazetteer:locality>Lago Tiaracá</Gazetteer:locality>
5. <Gazetteer:county>Novo Airão</Gazetteer:county>
6. <geosparql:hasGeometry rdf:resource="&Gazetter;/Geometry/713"/>
7. </owl:NamedIndividual>
8. <!– Gazetteer:Geometry/713 -->
9. <owl:NamedIndividual rdf:about="&Gazetteer;Geometry/713">
10. <geosparql:asWKT rdf:datatype="geow:wktLiteral">
11. <![CDATA[http://www.opengis.net/def/crs/OGC/1.3/CRS84 Point(0.20000000298023224 -66.0)]]>
12. </geosparql:asWKT>
13. </owl:NamedIndividual>
```

Fig. 6: Linking one location to a point through the GeoSPARQL ontology.

increased to 30.78% (around 20,000 records were inserted) and in SpeciesLink increased to 37.33% (around 30,000 records were inserted). It represents a significant increase of around 90% in the number of records with geographical information. Thus, we can affirm that the use of our Gazetteer (with the summarization technique) can lead to a significant increase in the geographical information in typical biodiversity data.

Another contribution refers to the reduction of inaccurate geographic information, shown in section 3. After applying the summarization technique, we can test, using the GeoSPARQL endpoint of the Gazetteer, for points outside polygons. Using a polygon that represents the Amazon State borders, we can delete any records with coordinates outside it. As a result 43,000 records with inaccurate locations were removed from the sample (those located at sea, in neighboring countries to Brazil or in other states), making the data more accurate.

We also tested a query sets against our Gazetteer and others SPARQL endpoints. For this experiment we used SPARQL/GeoSPARQL endpoints from the W3C[4] list of endpoints.

We began selecting three endpoints, DBpedia, Factor and *GeoSPARQL*. The Factor and *GeoSPARQL* endpoints both

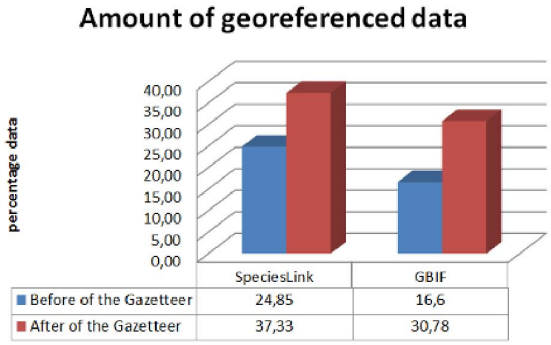


Fig. 7: Geographical information recover after the development of the Gazetteer.

contain information from Geonames, we then decided to drop *GeoSPARQL* from the experiment because it only has information about Brazilian municipalities, not covering places such as forests, reserves and lakes. Thus, we used the Factor endpoint, which contains information about Geonames and Wikipedia data, and the Dbpedia endpoint, which contains only information from Wikipedia.

To evaluate these endpoints, we selected a sample of 60

[4]http://www.w3.org/wiki/SparqlEndpoints

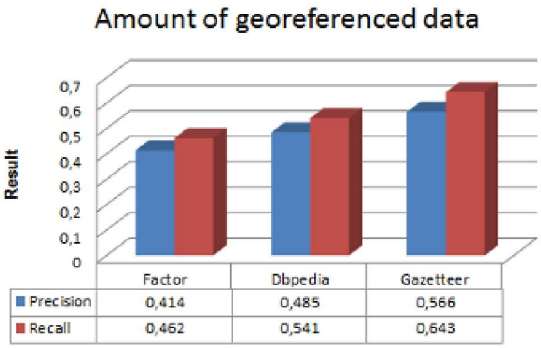


Fig. 8: Search using semantic web technologies

```
#Prefixes Hidden
SELECT ?p ?a ?w1 ?w2
WHERE {
?p rdf:type lgdo:Farm ;
   geo:hasGeometry ?g2 .
?g2 geo:asWKT ?w2 .
?a rdf:type dbp:ProtectedArea ;
   geo:hasGeometry ?g1 .
?g1 geo:asWKT ?w1
FILTER geof:sfWithin(?w2, ?w1)
}
```

Fig. 9: Search of Farms within a reserve

localities and created a database containing information relevant for all queries. The results of query sets used against Factor, DBpedia and our Gazetteer are show in Figure 8. Our Gazetteer obtained the best results for precision and recall, in relation to the other endpoints. This result was due to the fact that the other repositories do not contain place names relevant to the locations in the Amazon State where specimens were collected. They are not targeted to geographic information relevant to biodiversity data as our Gazetteer is. That is another motivation for the creation of another GeoSPARQL Gazetteer.

The use of a GeoSPARQL endpoint, to make the data from our Gazetteer public, enables the use of complex semantic queries searches. For instance, asking for farms that are near a forest reserve or records with coordinates outside the Amazonas State borders, as in the example shown in Figure 9. The inference necessary to solve these queries is possible, due to the fact that GeoSPARQL can reason about geometries and SPARQL logic integrating the two.

VI. Conclusion and Future Works

We presented the development of a gazetteer as a GeoSPARQL endpoint for biodiversity data from GBIF and SpeciesLink sites. We also demonstrate that this gazetteer can add absent geographic coordinates to biodiversity records and eliminate inaccurate geographic information.

We presented the problems faced during the gazetteer development and how we solve them. In addition, we made our Gazetteer openly available as a GeoSPARQL endpoint (at *http://biomac.icmc.usp.br:8088/parliament/*), a LOD resource that we expect will be very useful for the biodiversity community.

As future work, we intend to expand the Gazetteer to include data from other sources and regions of the world. We also plan to build interfaces to allow the curation of geographic data by users, to form a Collaborative Gazetteer, and a standard quality code for the geographic data provided, so users can have an idea of the data accuracy.

Acknowledgment

The authors would like to thank CAPES, a research foundation from the Brazilian federal government, for financing this work.

References

[1] S. Polasky, C. Costello, and A. Solow, "The Economics of Biodiversity," in *Handbook of Environmental Economics* (K. G. MÃ€ler and J. R. Vincent, eds.), vol. 3 of *Handbook of Environmental Economics*, ch. 29, pp. 1517–1560, Elsevier, June 2005.

[2] F. K. Amanqui, K. J. Serique, F. Lamping, A. C. F. Albuquerque, J. L. C. D. Santos, and D. A. Moreira, "Semantic search architecture for retrieving information in biodiversity repositories.," in *ONTOBRAS* (M. P. Bax, M. B. Almeida, and R. Wassermann, eds.), vol. 1041 of *CEUR Workshop Proceedings*, pp. 83–93, CEUR-WS.org, 2013.

[3] F. B. Gil, N. P. Kozievitch, and R. da Silva Torres, "Geonote: A web service for geographic data annotation in biodiversity information systems.," *JIDM*, vol. 2, no. 2, pp. 195–210, 2011.

[4] C. Yesson, P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, R. J. White, A. C. Jones, F. A. Bisby, and A. Culham, "How global is the global biodiversity information facility?," *PLoS ONE*, vol. 2, pp. e1124+, Nov. 2007.

[5] P. Lemes, F. Famv, G. Tessarolo, and R. D. Loyola, "Refinando dados espaciais para a conservacão da biodiversidade," *Natureza & Conservacão*, vol. 9, pp. 240–243, 2011.

[6] C. Kessler, K. Janowicz, and M. Bishr, "An agenda for the next generation gazetteer: Geographic information contribution and retrieval," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, (New York, NY, USA), pp. 91–100, ACM, 2009.

[7] K. Janowicz and C. KeBler, "The role of ontology in improving gazetteer interaction," *International Journal of Geographical Information Science*, vol. 22, no. 10, pp. 1129–1157, 2008.

[8] T. H. V. M. Moura and C. A. D. Jr., "Expansão do conteúdo de um gazetteer: nomes hidrográficos," in *Anais...* (L. M. Namikawa and V. Bogorny, eds.), (São José dos Campos), pp. 78–83, Simpósio Brasileiro de Geoinformática, 13. (GEOINFO), Instituto Nacional de Pesquisas Espaciais (INPE), 2012.

[9] C. Gouvea, S. Loh, L. F. F. Garcia, E. B. da Fonseca, and I. Wendt, "Discovering location indicators of toponyms from news to improve gazetteer-based geo-referencing.," in *GeoInfo* (M. T. M. de Carvalho, M. A. Casanova, M. Gattass, and L. Vinhas, eds.), pp. 51–62, INPE, 2008.

[10] J. L. C. dos Santos, *A biodiversity information system in an open data/metadatabase architecture*. PhD thesis, Enschede, 2003.

[11] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, and C. A. Davis, "Discovering geographic locations in web pages using urban addresses.," in *GIR* (R. Purves and C. Jones, eds.), pp. 31–36, ACM, 2007.

[12] J. A. Aslam, E. Pelekhov, and D. Rus, "The star clustering algorithm for static and dynamic information organization.," *J. Graph Algorithms Appl.*, vol. 8, pp. 95–129, 2004.

[13] Y. Yin and K. Yasuda, "Similarity coefficient methods applied to the cell formation problem: A taxonomy and review," *International Journal of Production Economics*, vol. 101, no. 2, pp. 329–352, 2006.

[14] R. Battle and D. Kolas, "Enabling the geospatial semantic web with parliament and geosparql.," *Semantic Web*, vol. 3, no. 4, pp. 355–370, 2012.