Proceedings XII GEOINFO, November 27-29, 2011, Campos do Jordão, Brazil. p 111-116.

# Using linked data to extract geo-knowledge

**Matheus Silva Mota**[1], **João Sávio Ceregatti Longo**[1]
**Daniel Cintra Cugler**[1], **Claudia Bauzer Medeiros**[1]

[1]Institute of Computing – UNICAMP
Campinas, SP – Brazil

{matheus,joaosavio}@lis.ic.unicamp.br, {danielcugler,cmbm}@ic.unicamp.br

***Abstract.*** *There are several approaches to extract geo-knowledge from documents and textual fields in databases. Most of them focus on detecting geographic evidence, from which the associated geographic location can be determined. This paper is based on a different premise – geo-knowledge can be extracted even from non-geographic evidence, taking advantage of the linked data paradigm. The paper gives an overview of our approach and presents two case studies to extract geo-knowledge from documents and databases in the biodiversity domain.*

## 1. Introduction

There is extensive research on extracting geographical knowledge from documents, mostly based on text analysis on documents and textual fields in databases. Basically, those papers try to find geographic references (e.g., matching place names according to a dictionary), and subsequently correlate the text with specific regions, points etc. [Odon de Alencar et al. 2010, Strötgen et al. 2010]. This often implies in issues of georeferencing, token indexing and document corpus analysis algorithms (e.g., using gazeteers [Nadeau et al. 2006] or geographical databases [Gomes and Medeiros 2007]). Furthermore, there are problems related to the heterogeneity of formats, since most of the solutions focus on interoperable formats (e.g., HTML) or some specific format (e.g., PDF).

Our work differs from such research efforts in three directions. (i) In case of documents, rather than analyzing the document itself, our strategy focus on an intermediate and interoperable document descriptor; (ii) instead of focusing only in geographical evidence, we also process other elements that may indirectly be associated with geo-information; (iii) we take advantage of the Open Linked Data Initiative to infer geo-knowledge.

This paper presents our work and two case studies involving our approach for the biodiversity domain. The first study processes research papers concerning biodiversity issues to extract geographic knowledge from non-geographic elements via linked data. The second case study processes a database of metadata of sound recordings of animals connecting the metadata with ontology terms and linked structured data. The first case study uses the notion of document descriptor and linked data, while the second connects metadata fields directly to semantic information.

This work is being developed at *LIS – Laboratory of Information Systems* – at the Institute of Computing, UNICAMP. This paper is organized as follows. Section 2 introduces concepts and related work. Section 3 presents two case studies and our effort to extract geo-knowledge from documents and databases using linked data. Finally, Section 4 presents conclusions and ongoing work.

## 2. Concepts and Related Work

### 2.1. Semantic Web, Linked Data and the Linking Open Data Project

The Semantic Web is commonly defined as the Web of Data [Auer et al. 2007]. The main difference between the Web as we know today and the Semantic Web is the focus on the meaning of the data, not only on availability and sharing as before. This information is not related to human consumption, but aims to help machines to understand and consume the information on the World Wide Web [Auer et al. 2007, Berners-Lee et al. 2001].

The notion of *Linked Data* appeared in the Semantic Web context. The term is related to a set of practices for publishing and sharing structured data on the Web. Basically, Linked Data uses the RDF (Resource Descriptor Framework) format to make typed statements that link things [Bizer et al. 2009b, Bizer et al. 2009a]. The 4 "rules" of linked data are: (i) Use URIs as names of things; (ii) use HTTP URIs so that people can look up those names; (iii) when someone looks up a URI, provide useful information; and (iv) include links to other URIs, so they can discover more things [Berners-Lee et al. 2001].

The Linking Open Data (LOD) is a W3C project related to the linked data publishing method. Its main goal is make several open data sets available and connected on the Web (such as DBPedia, Geonames, WordNet, the DBLP bibliography, the GeoSpecies Knowledge Base etc.). To do that, the data sets must publish the data as RDF, using URIs to link resources on/via Web [Auer et al. 2007, Bizer et al. 2009a]. Hence, LOD aims to create a machine consumable interlinked graph. The right part of Figure 1 shows some data sets that are interlinked by the project – 203 data sets, over 25 billion RDF triples interlinked by around 395 million RDF links (as of September 2010).

#### 2.1.1. Semantic Annotations for Geographical Data

The main idea of Semantic Annotations is derived from textual annotations [Oren et al. 2006]. Such annotations can have different objectives and be structured in many forms, e.g., free remarks, tags, floating layers [Euzenat 2002]. There are three main approaches to create annotations: manually [Lesaffre et al. 2003], semi-automatically [Macario and Medeiros 2009] and automatically [Cano 2004].

Annotations are used, among others, to describe a resource and what it represents. Informal ones are usually inserted on documents for human consumption. This hampers computer processing and annotation exchange. Semantic annotations appeared with the purpose of third-party interpretation, providing explicit and machine interpretable semantics, as supported by Semantic Web standards [Euzenat 2002]. Annotations acquire more semantics when they follow structural schemes and relate concepts and relationships between concepts and/or resources. This strategy allows machine consumption, therefore the development of new types of applications [Kiryakov et al. 2004], such as text categorization, content and multimodal information retrieval.

Semantic annotations in the geographic context should also consider the spatial component. Therefore, "the geospatial annotation process should be based on geospatial evidence - those that conduct to a geographic locality or phenomenon" [Macario and Medeiros 2009].

From a high level point of view, Wikipedia, for instance, has some kind of geographic semantic annotations. Latitude and longitude are provided through a coordinate template named *geotag*. Thus, the task to get this kind of information is straightforward. Relying on this fact, [Okamoto et al. 2010] focus on extracting and visualizing spatiotemporal data from Wikipedia. Moreover, [Odon de Alencar et al. 2010] show the feasibility of classifying Wikipedia documents according to their association to places. The method tries to find evidence of localization, searching in the graph formed by links. The authors claim it is almost always possible to say which location is related to a particular document.

### 2.2. Shadow-driven Document Representation

The SdR – *Shadow-driven Representation* – strategy is based on building an interoperable document descriptor that summarizes elements of a document. A *document shadow* can be seen as a generic structure (well formed XML) that isolates relevant elements of a document from its format, allowing its indexing and annotation/retrieval. These elements (e.g., title, authors, tables, figures, captions, sections) are defined by users: distinct group of users may have different needs on document management. The concept of shadow is akin to that of image descriptor: it summarizes key aspects of a document according to a predefined set of document features (elements of interest).

Figure 1 represents an abstraction of the SdR approach. A shadow element is associated to a fragment of a document (extreme left). The middle of Figure 1 (b) presents an abstraction of the internal structure of a shadow, with contents and structures: here, a document contains pages, which contain paragraphs etc. The production of a document shadow is divided in two steps: (a) Definition, by users. of the elements of interest that should be present in the shadow (the schema); and (b) instantiation of the shadow for each document, based on these elements.

## 3. Using linked data to extract geo-knowledge: Case Studies

This section presents two case studies where we use Linked Data to extract geo-knowledge from papers on biodiversity (Section 3.1) and from a metadata database (Section 3.2).

### 3.1. Extracting geo-knowledge from Biodiversity Documents Using Shadows

Rather than analyzing the text directly to extract geo-knowledge, the basis of our strategy focus on extracting this knowledge from shadows. This strategy, developed by us, isolates concerns on document format from query processing itself (and has allowed us to use a single set of code and algorithms to extract geographical knowledge from documents created using Adobe PDF, MS-Word and Open Office).

Instead of restricting ourselves to geographical data, we also process data indirectly associated with geographic references (e.g., images can be connected to their meaning on the semantic web, the authors of a paper can be connected to their birthplaces, conference proceedings can be connected with where the conference took place or the address of the publishers). The middle and right parts of Figure 1 present an abstraction of the connection between a shadow element (addressable via URI) and its meaning in a specific data set of the Linking Open Data Project (LOD).

For instance, consider a paper that contains an image of an animal. The image label identifies the species name. Using the species name, we can find its URI in the

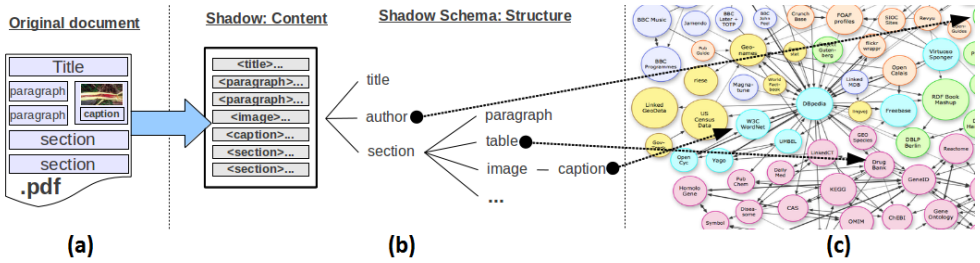


**Figure 1. Abstraction of the relation between a shadow, the corresponding document and the link between an element and an external data set**

LOD, and therefore additional geo-information about that species which are dispersed over different data sets. The paper's author can also be linked similarly, and so on. Hence, the shadow can now be used to answer queries such as "*what researchers have written papers on animals that are found within $X$ kilometers of their work place?*", or "*group papers by geographic regions of where the species described can be found*", or "*given a document, show where mentioned species can be found*", or even "*which documents mention species that appear in a polygon $P$?*".

### 3.2. Using Geo-knowledge and Linked Data to Improve Queries in the Biodiversity Context

Between 1961 and 2010, researchers from Fonoteca Neotropical "Jacques Vielliard" – UNICAMP – have recorded about 20.000 animal sounds, creating the largest collection of animal recordings in the Neotropics. When biologists record animal sounds, a set of metadata related to the recordings is also collected – e.g., air and/or water temperature, weather conditions, country/state/city where the sound was recorded, and so on.

Managing and retrieving animal sounds and their metadata pose countless challenges. We developed a web system where one can, among other functionalities, retrieve animal sounds based on their metadata [Cugler et al. 2011]. Since most of the data were collected in the 70's and 80's, coordinates are not provided. Moreover, even location names may not allow determining these coordinates, since they may contain notes such as "São Marcos Indian Reservation, Namun Kurá Village". One additional issue is that biologists would like to further exploit the metadata, by finding correlations with other facts.

To overcome these challenges, we performed a case study using this collection. Our focus was to use geo-knowledge and linked data as a bridge to connect animal recordings with the LOD data sets (see screen copy of our prototype, Figure 2). Metadata fields are used as concepts to be sought in LOD (in data sets such as DBPedia and GeoSpecies Knowledge Base). Once the concept is found in the LOD, it is possible to extract associated geo-knowledge. For instance, if a metadata field has a city name, then latitude/longitude can be retrieved from LOD, stored in a geographic database and used for other more complex queries. Linked data is able to provide additional information - e.g., region surface, altitude, description, population density, climate and time zone. Other kinds of information can also be retrieved, e.g., details about the animal whose sound was collected.

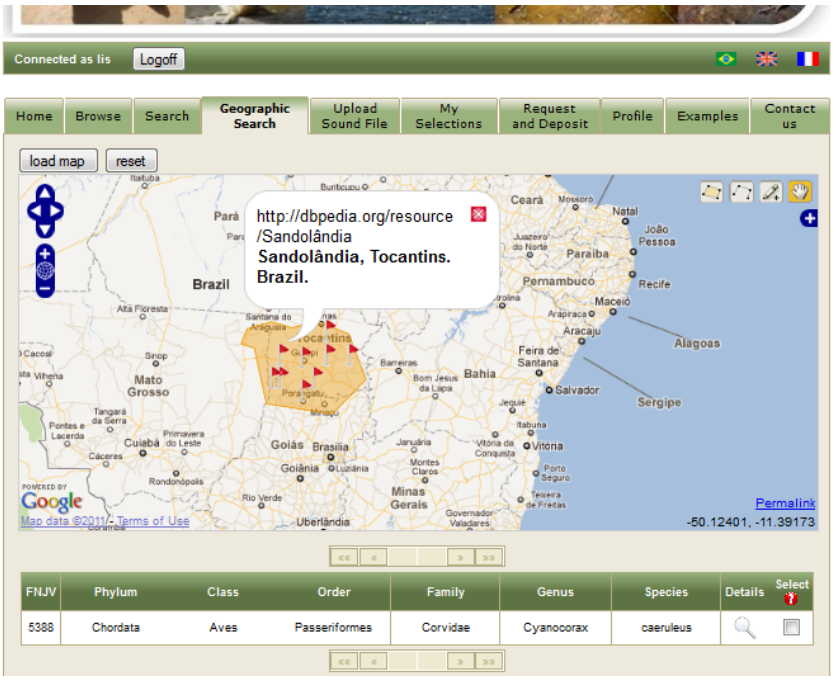


**Figure 2. Screen copy of the prototype**

The link between recordings and data from the LOD data sets provides plenty of semantic information. The limit of available fields for filtering queries is proportional to the size of the LOD Initiative. If it grows up, so do the filtering possibilities.

## 4. Concluding Remarks

This paper describes work in progress in combining shadows and semantic information via LOD. It proposes a different approach to extract geo-knowledge from documents and textual fields in databases. Our strategy adopts a notion of "document descriptor" (shadows) to handle the documents independently of file formats. It also takes advantage of the LOD project to extract geo-knowledge from non-geographic information.

We present two case studies where we connect elements of documents (through shadows) and a metadata database in biodiversity with open data sets (such as DBPedia and GeoSpecies Knowledge Base) in order to extract geo-knowledge and allow more sophisticated queries that will go beyond geographic references.

## Acknowledgments

Research partially financed by CNPq, FAPESP, CAPES and INCT in Web Science.

## References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer Berlin / Heidelberg.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):28–37.

Bizer, C., Heath, T., and Berners-Lee, T. (2009a). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009b). Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165.

Cano, P. (2004). Automatic sound annotation. In *In IEEE workshop on Machine Learning for Signal Processing*, pages 391–400.

Cugler, D., Medeiros, C., and Toledo, L. (2011). Managing animal sounds-some challenges and research directions. *Proceedings V eScience Workshop - XXXI Brazilian Computer Society Conference*.

Euzenat, J. (2002). Eight questions about semantic web annotations. *IEEE Intelligent S.*, 17(2):55–62.

Gomes, L. C. and Medeiros, C. B. (2007). Ecologically-aware queries for biodiversity research. In *Proceedings of GeoInfo - Brazilian Geoinformatics Symposium*, pages 73–84.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49 – 79.

Lesaffre, M., Tanghe, K., Martens, G., Moelants, D., Leman, M., Baets, B. D., Meyer, H. D., and Martens, J.-P. (2003). The mami query-by-voice experiment: Collecting and annotating vocal queries for music information retrieval. In *In: Proceedings of the International Conference on Music Information Retrieval*, pages 26–30.

Macario, C. G. N. and Medeiros, C. B. (2009). A framework for semantic annotation of geospatial data for agriculture. *Int. J. Metadata, Semantics and Ontology - Special Issue on "Agricultural Metadata and Semantics"*, 4:118–132.

Nadeau, D., Turney, P., and Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Advances in Artificial Intelligence*, volume 4013 of *Lecture Notes in Computer Science*, pages 266–277.

Odon de Alencar, R., Davis, Jr., C. A., and Gonçalves, M. A. (2010). Geographical classification of documents using evidence from wikipedia. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 12:1–12:8. ACM.

Okamoto, A., Yokoyama, S., Fukuta, N., and Ishikawa, H. (2010). Proposal of spatiotemporal data extraction and visualization system based on wikipedia for application to earth science. In *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*, pages 651–656.

Oren, E., Moller, K. H., Scerri, S., Handschuh, S., and Sintek, M. (2006). What are semantic annotations?? Technical report, DERI Galway.

Strötgen, J., Gertz, M., and Popov, P. (2010). Extraction and exploration of spatio-temporal information in documents. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 16:1–16:8, New York, NY, USA. ACM.