

# Inferring Geographical Ontologies from Multiple Resources for Geographical Information Retrieval

[Extended Abstract]

Davide Buscaldi  
Universidad Politécnica de  
Valencia  
Camino de Vera, s/n  
Valencia, Spain  
dbuscaldi@dsic.upv.es

Paolo Rosso  
Universidad Politécnica de  
Valencia  
Camino de Vera, s/n  
Valencia, Spain  
proso@dsic.upv.es

Piedachu Peris García  
Universidad Politécnica de  
Valencia  
Camino de Vera, s/n  
Valencia, Spain  
pperis@dsic.upv.es

## ABSTRACT

Most of the information available in electronic format, such as in the World Wide Web or in digital libraries, involves some kind of spatial awareness. For instance, news usually describe an event and the place where this event occurred: “Earthquake in Turkey”, “Visit of the Pope in Valencia”. Currently, the Information Retrieval (IR) research community is increasing its efforts dedicated to the retrieval of geographical information, as testified by the creation of the GeoCLEF<sup>1</sup> [5] evaluation exercise at the CLEF 2005, recently repeated in 2006, and the advances of the SPIRIT<sup>2</sup> project [6]. These efforts are aimed to the solution of typical issues of the geographical IR task. In many cases, explicit geographical information is missing from the documents, for instance the indication of a broader geographical entity is omitted when it is supposed to be well-known to the readers (e.g. usually *France* is not named in a news related to *Paris*). Another common problem is the synonymy, when there are many ways to indicate a geographical entity. This is particularly true for foreign names, where spelling variations are frequent. The solution to these problems has been generally individuated in the use of geographical-oriented ontologies [4, 6]. The manual construction of this kind of resources is usually a long, laborious process, and in many cases they are not freely available, such as the Getty Thesaurus of Geographical Names<sup>3</sup> (TGN). In order to overcome this issue, we made some attempts [2, 3] to use the geographical information included in WordNet, the well-known general domain ontology developed at the University of Princeton [7]. Unfortunately, the quantity of geographical information included in WordNet is quite small. Although it is quite difficult to calculate the number of geographi-

cal entities stored in WordNet, due to the lack of an explicit annotation of the synsets, we retrieved some figures by means the *has\_instance* relationship, resulting in 654 cities, 280 towns, 184 capitals and national capitals, 196 rivers, 44 lakes, 68 mountains. On the other hand, gazetteers like the Geonet Names Server<sup>4</sup> (GNS) and the Geographic Names Information System<sup>5</sup> (GNIS) are freely available and provide plenty of geographical informations. The problems of these resources is that they do not organize the information in a structured way like ontologies, and that they contain *too many* names; therefore, increasing the ambiguity of geographical names (for instance, 16 places named “*Genoa*” can be found in various locations all over the world: one in Italy, another in Australia and the remaining ones in the United States). Encyclopedias also contain a quantity of geographical information. One of the most interesting and recent phenomena in the Web is the success of Wikipedia<sup>6</sup> as source of information. The Wikipedia community itself is currently working in order to improve the quality of the geographical section (WikiProject Geography<sup>7</sup>). We already studied the possibility of using Wikipedia for Question Answering [1], a task related to the IR field, and we realized that it could be exploited also for the Geographical Information Retrieval, since the articles usually include useful information (such as boundaries) that can be used in order to extract relationships among geographical entities. Each of the discussed resources presents advantages and disadvantages. In this paper we describe our work in order to integrate the information extracted from WordNet, the GNS and GNIS gazetteers and Wikipedia into a geographical ontology. For instance, gazetteers lack informations about the composition of geo-political entities such as Europe, England, Scotland. This information can be retrieved by means of WordNet and/or Wikipedia. Our ontology has been implemented as a Prolog database that can be easily expanded with both new data and relationships. The first step was to read the GNS and GNIS data, extracting information of containment (region containing a city, states containing a region/county). The names extracted were passed through the set of geographical articles of a snapshot of the English edition of Wikipedia in order to filter out uncommon names

<sup>1</sup><http://ir.shef.ac.uk/geoclef/>

<sup>2</sup><http://www.geo-spirit.org>

<sup>3</sup><http://www.getty.edu>

<sup>4</sup><http://earth-info.nga.mil/gns/html/index.html>

<sup>5</sup><http://geonames.usgs.gov/domestic/index.html>

<sup>6</sup><http://www.wikipedia.org>

<sup>7</sup>[http://en.wikipedia.org/wiki/Wikipedia:WikiProject\\_Geography](http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geography)

(we made the assumption that well-known places are described at least once in the encyclopedia). We used the Xapian<sup>8</sup> search engine to index the Wikipedia snapshot. The geographical pages were selected using a list of trigger words picked automatically from WordNet. Subsequently, the names obtained from WordNet through the *part\_of* and *synonymy* relationships were included in the ontology, if not already present (a comparison of name and hierarchies was used for this step). Finally, state and region boundaries were extracted from Wikipedia using some shallow patterns (regular expressions) based on the structure of the articles. The resulting ontology can be used to improve the performance of our GeoCLEF Geographical IR system.

## Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

## General Terms

Geographical Information Retrieval, Ontologies, Wikipedia, WordNet, gazetteers

## 1. REFERENCES

- [1] D. Buscaldi and P. Rosso. Mining knowledge from wikipedia for the question answering task. In *Proceedings of the LREC 2006*, 2006.
- [2] D. Buscaldi, P. Rosso, and E. Sanchis. Using the wordnet ontology in the geoclef geographical information retrieval task. In *Proceedings of the CLEF 2005*, 2005.
- [3] D. Buscaldi, P. Rosso, and E. Sanchis. Wordnet as a geographical information resource. In *Proceedings of the 3rd Global WordNet Association (GWA06)*, 2006.
- [4] G. Fu, C. Jones, and A. Abdelmoty. Building a geographical ontology for intelligent spatial search on the web. In *Proceedings of the IASTED International Conference on Databases and Applications*, 2005.
- [5] F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough. Geoclef: the clef 2005 cross-language geographic information retrieval track. In *CLEF 2005 Working Notes*, C.Peters Ed., 2005.
- [6] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies: An overview of the spirit project. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, 2002.
- [7] G. A. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41, 1995.

---

<sup>8</sup><http://xapian.sourceforge.net>